LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN
numero 180
agosto 1983
anno XVI
volume XXXI

Un vettore che introduce nuovi geni nelle piante

Un batterio induce un tumore nelle piante ricorrendo a una forma naturale di ingegneria genetica. Il frammento di DNA che inocula può modificare geneticamente anche piante d'importanza economica

di Mary-Dell Chilton

Da molto tempo, forse da milioni di anni, il comune batterio del terreno *Agrobacterium tumefaciens* sta facendo quello che i biologi molecolari si stanno solo ora impegnando a fare. Esso inserisce, infatti, geni estranei nelle piante e le induce a esprimere quei geni sotto forma di proteine. Nel corso del processo costringe le cellule vegetali a proliferare e a formare una galla o tumore, soprattutto in corrispondenza della giunzione tra radice e fusto (colletto), o vicino a essa.

La galla, o tumore, del colletto è una malattia che può colpire una vasta gamma di dicotiledoni e provoca considerevoli perdite in certe colture, in particolare viti, piante con frutti provvisti di nocciolo (drupe) e piante ornamentali. Fin dall'inizio del secolo è stata analizzata dai fitopatologi per il suo impatto sull'economia e, negli ultimissimi anni, questo processo infettivo è risultato chiaramente una forma naturale di ingegneria genetica. *A. tumefaciens* e il tumore che genera sono diventati l'argomento di un'indagine estremamente competitiva da parte dei biologi molecolari di laboratori accademici e industriali. È probabile che quello che, per molto tempo, è stato considerato semplicemente come l'agente di una fastidiosa malattia delle piante diventerà uno strumento importante per la manipolazione genetica delle specie vegetali: precisamente per introdurre nuovi geni nelle piante e pertanto dare origine a nuove varietà con caratteri desiderati.

Nel 1907 Erwin F. Smith e C. O. Townsend dello US Department of Agriculture hanno identificato in *A. tumefaciens* l'agente causale del tumore del colletto: ciò è stato possibile isolando il batterio dalle galle e mostrando successivamente che la sua inoculazione in una ferita di una pianta sana avrebbe generato un nuovo tumore. In seguito è emerso chiaramente che l'infezione ha luogo solo in tessuti lesi di recente; la ferita diventa refrattaria solo dopo un paio di giorni e, di fatto, eliminando il «succo» che sgorga dalla ferita fresca si riesce a bloccare l'induzione del tumore. Non tutti i ceppi di *A. tumefaciens* sono virulenti, o capaci di indurre un tumore. Quelli virulenti sono in grado di legarsi a costituenti pectinici specifici della parete delle cellule vegetali. Le cellule batteriche penetrano in effetti solo in cellule vegetali lese e morte e stimolano a questo punto le cellule adiacenti, così da far loro formare un tumore. I batteri stessi, poi, proliferano in mezzo alle cellule tumorali.

Nel 1947, Armin C. Braun del Rockefeller Institute for Medical Research è riuscito per primo a coltivare in laboratorio, su un terreno completamente definito e contenente soltanto saccarosio e sali inorganici, il tessuto della galla, privo del batterio inducente. Ha trovato così che quel tessuto aveva la proprietà caratteristica di un tumore: una crescita incontrollata. Mentre le cellule o i tessuti vegetali normali, in coltura, crescono lentamente e solo in presenza di ormoni vegetali (citochinine e auxine), il tessuto della galla cresceva rapidamente anche in assenza di un qualsiasi ormone esogeno. E continuava a crescere a ritmo molto intenso anche quando veniva trasferito, di tempo in tempo, su terreno di coltura fresco. Braun ha dunque concluso che le cellule vegetali dovevano essersi in qualche modo trasformate (modificate) in cellule tumorali per l'azione di disturbo di *A. tumefaciens*. Egli ha anche avanzato l'idea che il batterio introducesse nella cellula vegetale «un principio in grado di indurre un tumore».

Un'importante chiave metabolica per riuscire a capire la natura del suddetto principio è stata scoperta negli anni sessanta da Georges M. Morel e collaboratori all'Institut National de la Recherche Agronomique di Versailles. Essi hanno trovato che le cellule del tumore del colletto sintetizzano una classe di composti chimici nuovi (da loro chiamati «opine»), i quali sono assenti nelle cellule normali della stessa pianta. Le opine sono derivati di comuni composti intermedi del metabolismo, per la maggior parte amminoacidi (i costituenti delle proteine) e vari chetoacidi e zuccheri. Due opine sono state studiate in modo approfondito: l'octopina, un composto dell'amminoacido arginina e dell'acido piruvico, e la nopalina, un prodotto dell'arginina e dell'acido alfa-chetoglutarico. Il gruppo di Morel ha anche potuto precisare due punti importanti. In primo luogo che l'eventuale sintesi da parte del tumore di octopina e opalina non dipende dalla specie della pianta ospite, ma dal ceppo del batterio che induce il tumore. In secondo luogo che un determinato ceppo batterico può svilupparsi a spese sia dell'octopina sia della nopalina, ma non di entrambe; tale ceppo può metabolizzare qualunque composto venga sintetizzato dal tumore che induce. Morel ha poi fatto una sorprendente previsione: ha avanzato l'idea che il batterio inserisca nella cellula vege-

tale un gene che controlla la sintesi dell'octopina o della nopalina; il principio che induce il tumore deve essere DNA.

Fu subito chiaro che, se le cose stanno in questi termini, il DNA deve trovarsi su qualche elemento mobile. Allen Kerr dell'Università di Adelaide ha notato che, inoculando due ceppi nella stessa pianta, uno virulento e l'altro no, la proprietà della virulenza veniva trasmessa dall'uno all'altro. Questo trasferimento aveva luogo anche tra specie di *Agrobacterium* molto differenti tra loro, rendendo così poco probabile il fatto che i geni responsabili della virulenza si trovassero sul cromosoma batterico. Kerr pensava che, con molta probabilità, essi erano trasportati da un virus batterico o da un plasmidio: un piccolo anello di DNA a duplice filamento, non cromosomico, che porta di solito geni accessori, ad esempio quelli responsabili della patogenicità, del metabolismo di composti rari o della resistenza agli antibiotici.

Ulteriori prove che la virulenza in *Agrobacterium* potesse essere determinata da un virus o da un plasmidio emersero dalla scoperta che la virulenza poteva andare perduta. Nel 1971, Robert H. Hamilton, Jr., e M. Z. Fall della Pennsylvania State University hanno scoperto un ceppo di *A. tumefaciens,* la cui virulenza è sensibile alla temperatura. Quando il ceppo *C*58 viene fatto crescere a una temperatura moderatamente elevata (37 gradi centigradi), perde la virulenza e non riacquista mai spontaneamente la capacità di indurre tumori. In altre parole, qualche cosa va definitivamente perduto da parte delle cellule batteriche di quel ceppo. Questo fatto implica nuovamente che i geni responsabili dell'induzione del tumore devono essere localizzati su un virus o su un plasmidio.

Questo DNA evasivo è stato identificato nel 1974. Jeff Schell e Marc Van Montagu assieme ai loro collaboratori dell'Università statale di Gand (Belgio) hanno trovato grossissimi plasmidi in tutti i ceppi virulenti; non erano invece necessariamente presenti plasmidi nei ceppi non virulenti. Si poté presto dimostrare che la perdita della virulenza da parte del ceppo *C*58 a 37 gradi centigradi è dovuta alla perdita del corrispondente plasmidio, il quale è l'elemento mobile di Kerr, che viene trasferito da un ceppo virulento a uno che non lo è mediante coniugazione, una versione batterica dell'accoppiamento. Esperimenti di coniugazione, effettuati da Rob A. Schilperoort e collaboratori dell'Università statale di Leida, hanno messo in evidenza che è il plasmidio a conferire al batterio la capacità di catabolizzare (demolire) l'octopina o la nopalina e di accrescersi così a spese di uno di questi composti, il che determina se il tumore sintetizzerà octopina o nopalina. Anche la diffusione di un particolare plasmidio responsabile della virulenza in una popolazione di *A. tumefaciens* è promossa dalla presenza dell'opina, la cui sintesi e il cui catabolismo sono da esso specificati: il trasferimento del plasmidio nel corso della coniugazione è specificamente indotto da una particolare opina.

In altre parole, il plasmidio *Ti* («*tumor inducing*») è l'elemento centrale di una interazione ecologica meravigliosa, estremamente evoluta. La presenza di un plasmidio *Ti* abilita un batterio a sovvertire il meccanismo metabolico delle cellule vegetali in modo che producano una sostanza che è inutile per la pianta, ma a spese della quale i batteri (e pertanto anche il plasmidio loro simbionte) possono proliferare. Lo sviluppo di un tumore fa sì che tale sostanza venga prodotta in notevole quantità. In natura il tessuto tumorale viene a trovarsi a livello del suolo (e presumibilmente le opine che produce si diffondono proprio qui): il trasferimento di plasmidi a batteri fino a quel momento non virulenti promuove la sopravvivenza sia delle specie batteriche sia dei plasmidi.

Se una galla del colletto viene prodotta mediante trasferimento genico e se i geni sono localizzati sul plasmidio *Ti,* chiaramente nelle galle del colletto dovrebbe essere presente DNA plasmidiale. All'Università di Washington a Seattle, Eugene W. Nester, Milton P. Gordon e io stesso, assieme a un numero di studenti che si preparavano alla laurea e al dottorato, abbiamo intrapreso la ricerca di quel DNA. A quell'epoca il modo migliore per individuare un particolare DNA (nel nostro caso il plasmidio *Ti*) in presenza di una notevole quantità di altro DNA (il DNA della pianta) consisteva nel misurare il tasso di riassociazione del DNA. Occorreva cioè denaturare la doppia elica di DNA (venivano separati i due filamenti complementari) mediante riscaldamento a 100 gradi centigradi; quando la temperatura veniva abbassata a 68 gradi i filamenti complementari che «si ritrovavano» si riassociavano, cioè subivano un processo di rinaturazione, formando di nuovo una doppia elica. Il tasso di riassociazione è proporzionale al quadrato della concentrazione del DNA. La nostra procedura è consistita nel confrontare il tasso di riassociazione del solo DNA denaturato del plasmidio *Ti* con il tasso di riassociazione dello stesso DNA in presenza di DNA vegetale normale, o in presenza di DNA del tumore del colletto. Se il DNA delle cellule tumorali includeva copie del DNA plasmidiale, si sarebbe aggiunto alla concentrazione dei filamenti complementari e, pertanto, avrebbe fatto subire al DNA plasmidiale (che potevamo individuare, in quanto era marcato con un isotopo radioattivo) un processo di rinaturazione più rapido, in una maniera quantitativamente prevedibile.

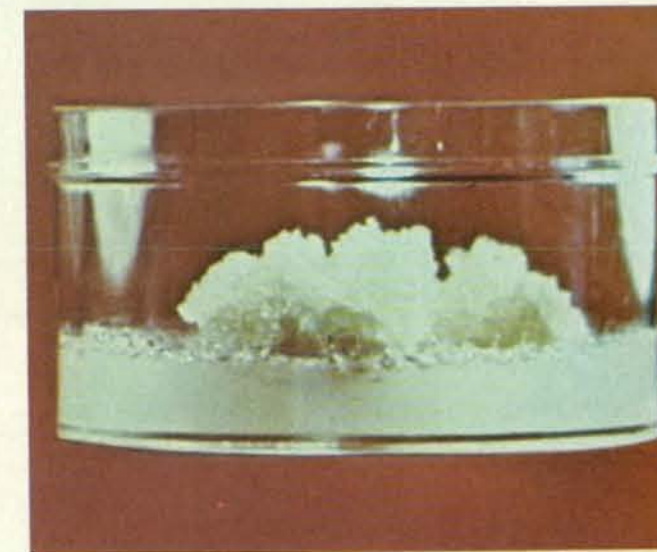
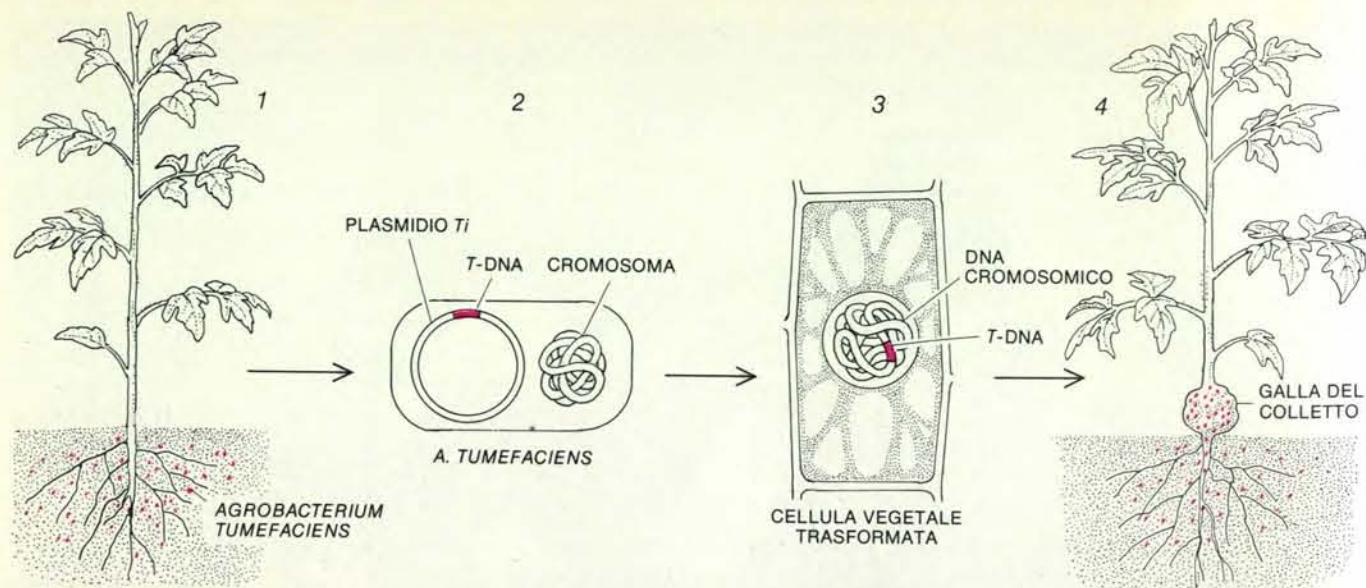
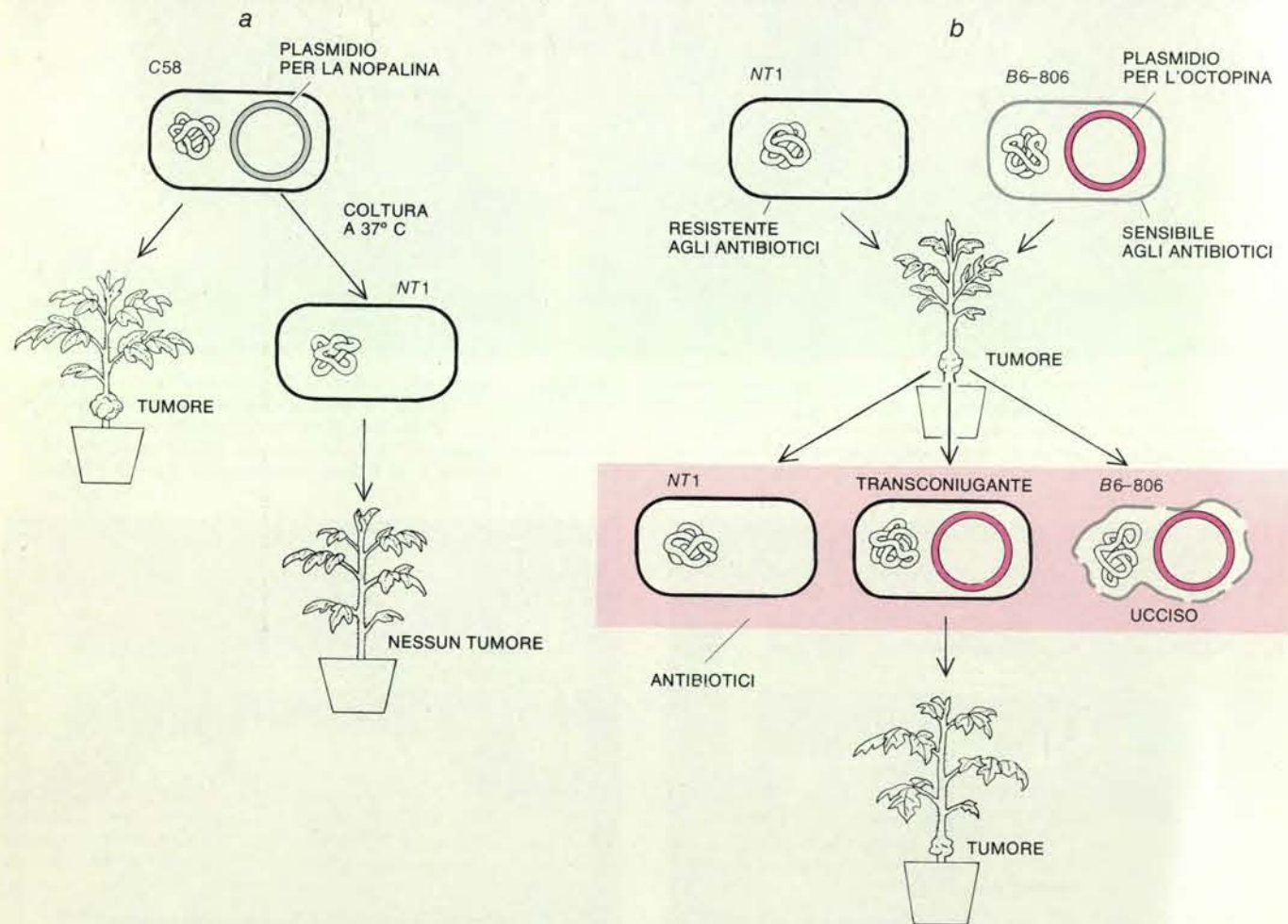
I nostri risultati iniziali hanno dimostrato che l'aggiunta di DNA denaturato del tumore a DNA denaturato del plasmidio non porta ad alcuna differenza significativa nel tasso di riassociazione, il che implicava chiaramente che nel tumore non c'era alcuna copia dell'intero plasmidio *Ti.* Avrebbe potuto essercene solo una parte? Abbiamo tagliato il DNA del plasmidio con una endonucleasi di restrizione, un membro di una famiglia di enzimi, ciascuno dei quali riconosce una particolare breve sequenza di nucleotidi (subunità di DNA) come unica sede dove possa aver luogo la scissione, e pertanto taglia sempre un certo DNA in una serie riproducibile di segmenti. L'endonucleasi *Sma* I taglia un plasmidio *Ti* in una ventina di segmenti di differenti dimensioni, cosicché la digestione di una certa quantità di DNA plasmidiale produce una ventina di tipi unici di segmenti, che possono essere separati mediante elettroforesi su gel in base alle loro dimensioni. Ogni gruppo di segmenti forma una stretta banda orizzontale su gel; questa può essere visualizzata mediante colorazione con una sostanza fluorescente. Il DNA di ogni banda può poi essere purificato per ulteriori esperimenti.

Quando abbiamo sottoposto a elettroforesi il DNA del plasmidio *Ti* e abbiamo quindi realizzato l'esperimento di rinaturazione con segmenti singoli di DNA, invece che con il plasmidio intero, ci siamo elettrizzati nel trovare che due dei segmenti subivano, in presenza del DNA tumorale, un processo di rinaturazione più veloce; gli altri segmenti no. Quei due segmenti del plasmidio *Ti* devono pertanto essere presenti nelle cellule tumorali. Successivi esperimenti effettuati con molte linee tumorali hanno mostrato che segmenti specifici del plasmidio *Ti,* in grado di indurre il tumore, sono sempre presenti nelle cellule della galla del colletto. Noi chiamiamo questi segmenti «DNA trasportato», o *T*-DNA.

L'esame dei tassi di rinaturazione è stato oggi sostituito da un metodo più diretto per individuare particolari segmenti di DNA inclusi in una grande quantità di DNA non specifico. L'analisi delle macchie, o metodo di Southern, messa a punto nel 1975 da E. M. Southern dell'Università di Edimburgo, combina l'elettroforesi con la rinaturazione ed è diventata uno strumento fondamentale d'indagine in genetica molecolare. Per trovare il DNA plasmidiale in cellule tumorali, si comincia a digerire il DNA delle cellule tumorali ad esempio con *Sma* I. Mentre una simile operazione porta a venti segmenti per ogni plasmidio, produce un milione circa di segmenti per il genoma (corredo genico complessivo) della cellula tumorale vegetale, che è enormemente più grande. Quando questi ultimi segmenti vengono dispersi su un gel mediante elettroforesi, formano quello che in apparenza sembra uno striscio continuo (anche se effettivamente si tratta di una serie di un milione circa di bande distinte). Qualunque frammento di *T*-DNA presente nel tumore resta incluso in quello striscio; la tecnica di Southern riesce a prelevarlo.

I segmenti di DNA tumorale vengono denaturati e quindi trasferiti su un foglio di carta da filtro alla nitrocellulosa mediante assorbimento; il foglio porta così un'impronta netta, ma invisibile del quadro originario. Il DNA si fissa alla carta, la quale viene incubata con una soluzione di segmenti denaturati e marcati con isotopi radioattivi di DNA plasmidiale, che

Su questa pianta di tabacco si è sviluppata una galla del colletto dove la parte superiore è stata asportata e il restante fusto sezionato è stato inoculato con *Agrobacterium tumefaciens.* I batteri infettanti contengono un DNA circolare accessorio: il plasmidio *Ti* (da *tumor inducing*). Un segmento di questo plasmidio, o *T*-DNA (DNA trasportato), è inserito nelle cellule e dà origine al tumore, in questo caso un teratoma, da cui, in modo disorganizzato, si sviluppano sia germogli sia radici. Si spera di sfruttare il plasmidio come vettore per introdurre geni nelle piante, così da ottenere piante migliorate. La fotografia è di Robert e B. Gillian Turgeon della Cornell University.

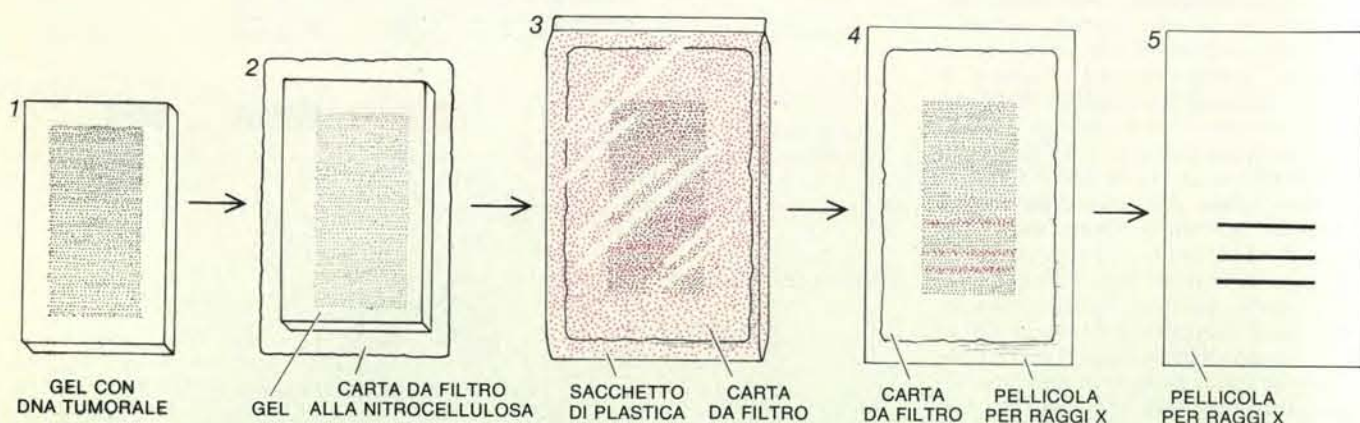
Nel laboratorio dell'autore alla Washington University di St. Louis è stato fatto crescere questo tessuto tumorale di tabacco. I batteri sono stati eliminati dalle linee cellulari derivate da piante infette, mentre la loro influenza sotto forma di cambiamento genetico rimane nelle cellule trasformate. Le cellule si sviluppano in coltura senza che siano state aggiunte auxine o citochinine, ormoni necessari per un normale sviluppo in vitro. Compaiono qui due tipi di crescita tumorale: un teratoma (*a sinistra*) e un callo o massa di cellule indifferenziate (*a destra*).

Il tumore comincia a formarsi nella pianta, quando i batteri penetrano in una zona lesa, in genere vicino al colletto, e si attaccano alle cellule *(1)*. Un batterio virulento contiene, oltre al DNA cromosomico, un plasmidio *Ti (2)*. Il *T*-DNA del plasmidio viene introdotto in una cellula e si integra con il suo DNA cromosomico *(3)*. Le cellule trasformate proliferano formando una galla *(4)*. Le cellule tumorali sintetizzano composti chiamati opine, che servono come sostanze nutritive alle cellule di *A. tumefaciens,* incluse nel tumore. Due opine ben studiate sono l'octopina e la nopalina. Il plasmidio *Ti* contenuto in un determinato ceppo del batterio induce la sintesi di una determinata opina.

Il plasmidio viene identificato come agente tumorigeno grazie a esperimenti come quelli qui illustrati. Il ceppo di *A. tumefaciens* indicato *C*58 è normalmente virulento *(a)*. Si sviluppa a spese della nopalina e induce tumori che sintetizzano questa sostanza. Dopo essersi sviluppato a 37 gradi centigradi, il ceppo perde la sua virulenza e la capacità di crescere a spese della nopalina; il plasmidio *Ti* che gli appartiene manca nel derivato non virulento *NT*1. La virulenza può essere restaurata nel derivato non virulento *NT*1 mediante appaiamento con un ceppo virulento *(b)*. Le cellule *NT*1 (mutanti resistenti ai due antibiotici) vengono inoculate in una pianta di pomodoro con *B*6-806, un ceppo virulento che si sviluppa a spese dell'octopina. Si forma un tumore, che viene triturato e successivamente sparpagliato in un terreno colturale assieme ai due antibiotici, in modo che le cellule *NT*1 resistenti formino delle colonie. Molte di queste cellule sono dei transconiuganti, cioè contengono il plasmidio per l'octopina del ceppo *B*6-806, che hanno acquisito attraverso la coniugazione, una versione batterica dell'accoppiamento. Queste cellule sono virulente, si sviluppano a spese dell'octopina e inducono tumori i quali sintetizzano la suddetta sostanza.

funge da sonda. Se nel tumore è presente DNA plasmidiale, il DNA che funge da sonda si riassocia con esso, formando DNA ibrido a duplice filamento; il resto della sonda viene eliminato, lasciando il DNA plasmidiale radioattivo legato ai filamenti complementari del DNA tumorale. La carta viene quindi messa a contatto con una pellicola fotografica per raggi X, la quale, quando viene sviluppata, mostra sull'autoradiogramma alcune bande scure nel punto in cui il *T*-DNA è stato incorporato nel DNA della pianta. Un confronto tra queste bande e le bande che compaiono su un gel contenente DNA plasmidiale puro digerito da *Sma* I mostra quali segmenti prodotti da questa endonucleasi si trovano, in tutto o in parte, nel *T*-DNA: si tratta dei segmenti 3*b*, 10*c*, 16 e 17.

Per capire nei particolari la struttura dei plasmidi *Ti* e del loro *T*-DNA è stato necessario in primo luogo determinare in che modo i segmenti plasmidiali si dispongono in un anello chiuso e quindi localizzare, sulla mappa circolare dei vari segmenti, i geni per particolari funzioni. Il nostro gruppo all'Università di Washington ha dedotto la prima mappa dei vari segmenti di un plasmidio *Ti* (il plasmidio responsabile della sintesi dell'octopina e designato come *B*6-806), digerendo il plasmidio con l'endonucleasi *Hpa* I e osservando in che modo i segmenti risultanti combaciavano con i segmenti dello stesso plasmidio, prodotti dall'endonucleasi *Sma* I. Con la tecnica di Southern abbiamo ottenuto un grande tracciato elettroforetico di bande di segmenti prodotti dalla digestione con *Hpa* I *(si veda l'illustrazione in questa pagina)* e l'abbiamo poi tagliato in tante strisce longitudinali. Un diverso segmento marcato, ottenuto per azione dell'endonucleasi *Sma* I, è stato fatto ibridare con ogni banda. Un segmento marcato che si lega a una sola banda *Hpa* I (banda corrispondente a un segmento ottenuto per digestione con l'endonucleasi *Hpa* I) deve collocarsi, nella costruzione della mappa, all'interno di quella banda; un segmento che si lega a due bande *Hpa* I deve coprire la giunzione tra queste due bande e così via. Applicando criteri come questi e analizzando i segmenti generati con parecchie endonucleasi, si riesce a costruire una mappa completa di segmenti.

La fase successiva è consistita nel ricostruire la posizione dei geni plasmidiali associando mutazioni in corrispondenza di particolari siti con variazioni di particolari funzioni. A questo scopo, diversi gruppi hanno sfruttato la tecnica della mutagenesi del trasposone. Alcuni geni responsabili della resistenza a un antibiotico sono portati da trasposoni, corti segmenti di DNA che tendono a saltare, in modo praticamente casuale, in nuove localizzazioni del cromosoma batterico o di un plasmidio. La presenza di un trasposone genera simultaneamente una mutazione (dato che l'inserimento del trasposone interrompe un gene sul cromosoma o sul plasmidio), identifica se stessa (in quanto il batterio è reso resistente a un antibiotico) e segnala la sua localizzazione (si modifica infatti il quadro dei segmenti ottenuti con l'endonucleasi a seguito dell'inserimento di nuovo DNA in un vecchio segmento, il che dà origine a nuovi segmenti). All'Università di Washington, Nester e David Garfinkel hanno introdotto il trasposone mediante un «vettore suicida»: un plasmidio ottenuto da John Beringer del John Innes Institute in Inghilterra, che porta il trasposone *Tn*5, responsabile della resistenza alla canamicina. Introdotto in *Agrobacterium,* questo plasmidio opportunamente si autodistrugge, ma solo dopo che il trasposone *Tn*5 ha avuto l'opportunità di saltare in un nuovo punto del cromosoma batterico o di un plasmidio *Ti.*

La mappa a vari segmenti, per questo plasmidio, deriva dalla sovrapposizione di due serie di segmenti ottenuti grazie a una endonucleasi. Con il metodo di Southern è stato realizzato un ampio tracciato in cui si distinguono i vari segmenti prodotti dalla scissione del plasmidio *Ti* con l'endonucleasi *Hpa* I *(a)*. Il foglio di carta da filtro viene quindi tagliato in strisce e ogni striscia viene incubata con una differente sonda di DNA, ottenuta per digestione con lo *Sma* I e marcata con un isotopo radioattivo. Questa sonda si riassocia con la sua controparte ottenuta per azione dell'*Hpa* I *(b)*. Per esempio, il segmento 4 ottenuto per azione dell'enzima *Sma* I si ibrida con i segmenti 2 e 4 ottenuti per azione di *Hpa* I. Il segmento 6 ottenuto con *Sma* I si ibrida solo con il segmento 2 ottenuto con *Hpa* I. Tutti questi dati forniscono un andamento lineare dei segmenti che si sovrappongono *(c)* e daranno alla fine una mappa completa del plasmidio circolare *Ti (d)*.

Il plasmidio *Ti* viene analizzato mediante elettroforesi su gel. Esso è sottoposto a digestione da parte di una endonucleasi di restrizione, enzima che scinde il DNA in punti specifici. I segmenti che sono prodotti dall'azione dell'endonucleasi *Sma* I (migliaia di copie di ogni segmento da migliaia di copie del plasmidio) sono posti in una fenditura a una estremità della striscia di gel di agarosio. Sottoposti a un campo elettrico, i segmenti di DNA dotati di carica elettrica negativa si spostano verso il polo positivo. Quanto minori sono le loro dimensioni, tanto più rapidamente essi si muovono, e formano così una serie di bande, che si possono mettere in evidenza ricorrendo a un colorante fluorescente che reagisce con il DNA. Un dato plasmidio, sottoposto a digestione con una data endonucleasi, dà luogo a una serie di bande caratteristiche.

La presenza di DNA plasmidiale (*T*-DNA) nel tessuto della galla viene dimostrata con il metodo di Southern. Segmenti di DNA tumorale, generati dalla digestione con *Sma* I, vengono separati mediante elettroforesi su gel. Le migliaia di bande di segmenti formano sul gel uno striscio continuo *(1)*. Il DNA è denaturato (i due filamenti vengono separati), viene trasferito su carta da filtro mediante assorbimento e quindi fissato su di essa *(2)*. La carta viene quindi incubata con DNA che funge da sonda *(in colore)*: il DNA del plasmidio *Ti*, o segmenti di esso, marcato con un isotopo radioattivo e denaturato *(3)*. Segmenti di DNA sonda si riassociano con DNA complementare, cioè «trovano» e si legano a bande costituenti il *T*-DNA. Il DNA sonda non legato è quindi eliminato *(4)*. Le bande di *T*-DNA sono rese visibili dall'autoradiografia *(5)*. I segmenti sono identificabili confrontandoli con il gel originario che contiene i prodotti della digestione da *Sma* I.

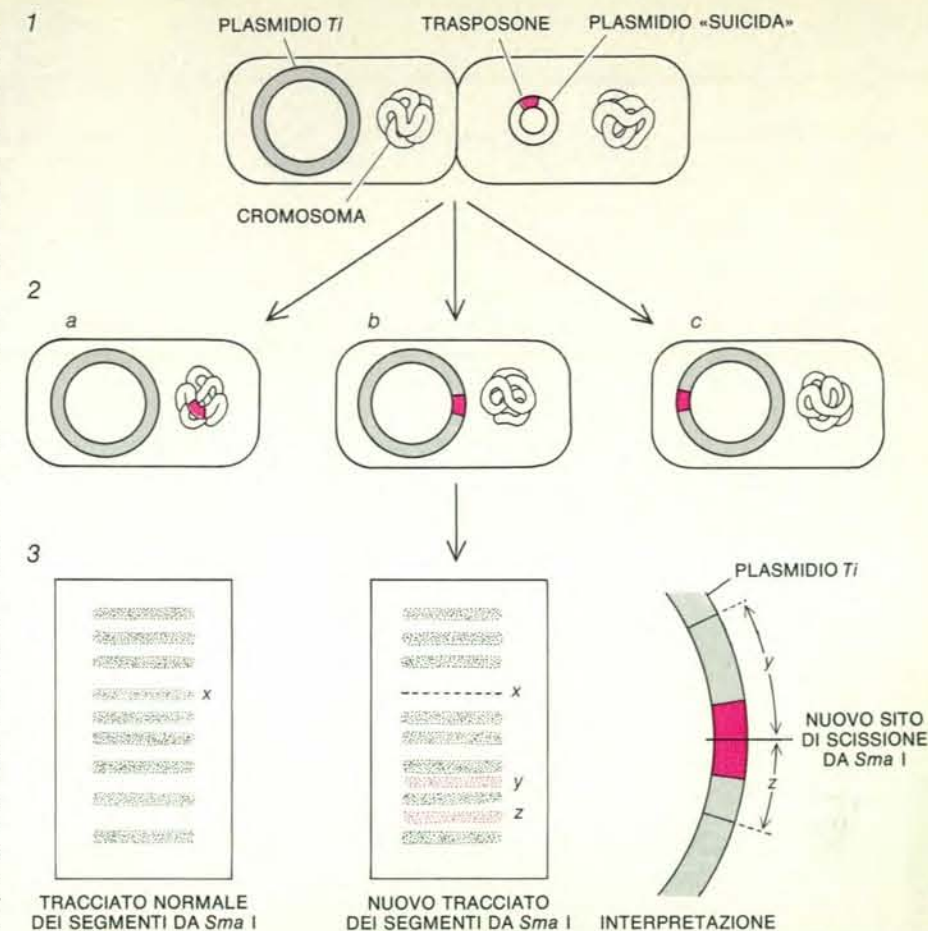
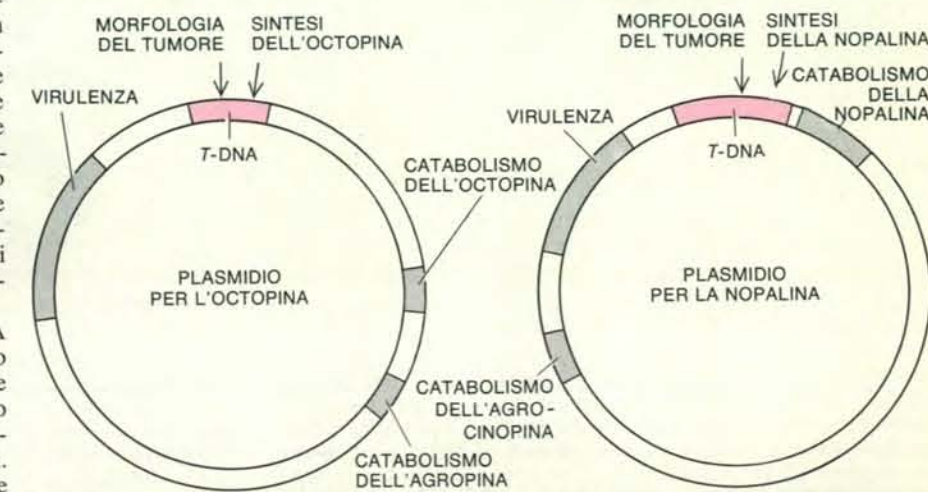
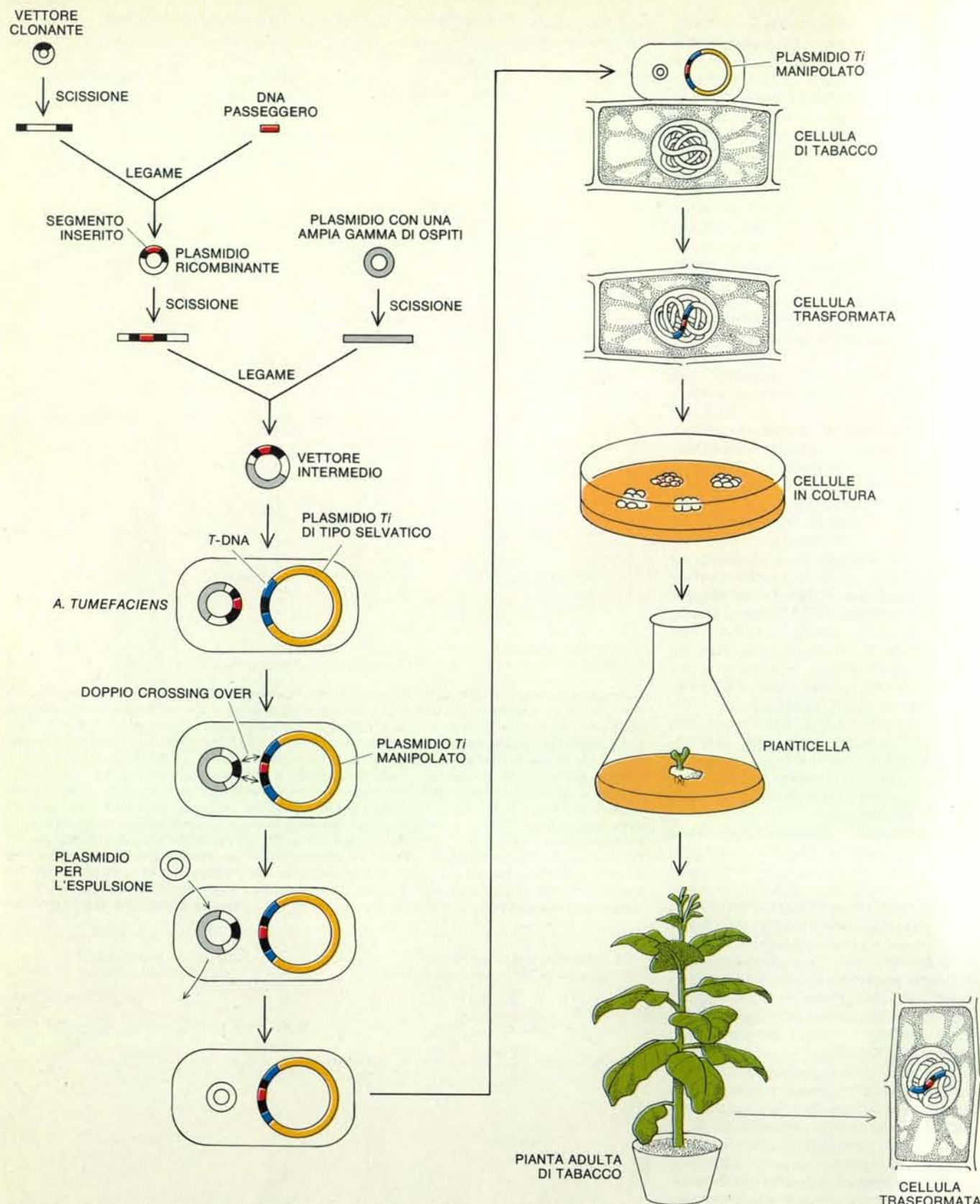
Si selezionano colonie batteriche resistenti alla canamicina, in cui il trasposone *Tn5* si trovi sul plasmidio piuttosto che sui cromosomi *(si veda l'illustrazione in alto nella pagina a fronte)*, e quindi le si analizza per rilevare qualsiasi cambiamento relativo alle proprietà genetiche. Molte mutazioni sono «silenti», ma in alcuni casi si osserva un cambiamento ben individuabile. Per esempio, se i batteri non riescono a catabolizzare l'octopina, ciò vuol dire che un gene necessario per questo processo deve aver subito una mutazione. Per poter stabilire la localizzazione di questo gene si isola il DNA del plasmidio *Ti* dai batteri mutanti, lo si digerisce con l'endonucleasi *Sma* I e si esamina il risultante tracciato elettroforetico, che è diverso dal normale: uno dei soliti segmenti (lo si chiami *X*) manca e sono invece visibili due nuovi segmenti; combinando le loro dimensioni si ottiene la dimensione del segmento *X* più *Tn5*. Il gene deve trovarsi all'interno di *X*. Sapendo che l'endonucleasi *Sma* I taglia il *Tn5* circa alla metà, si può determinare, partendo dalla dimensione dei due nuovi segmenti e da certe altre informazioni, la localizzazione precisa del gene.

Con questo tipo di analisi, condotto dal gruppo di Nester, da Schell e Van Montagu e dal gruppo di Schilperoort, si ottiene alla fine una mappa genetica per i plasmidi *Ti*, responsabili della sintesi dell'octopina e della nopalina. Un ampio settore del plasmidio alla sinistra del *T*-DNA porta le funzioni relative alla virulenza. In esso le mutazioni presumibilmente bloccano l'attacco del batterio a una cellula vegetale, il trasferimento del DNA del plasmidio in questa cellula o l'integrazione di questo DNA nel cromosoma della cellula vegetale, o forse anche parecchie di queste funzioni. Un altro ampio settore, sull'altro lato del *T*-DNA, controlla l'assunzione e il catabolismo dell'octopina o della nopalina. Sono stati anche localizzati in una mappa i geni che controllano la capacità di trasferimento del plasmidio durante la coniugazione, così come è avvenuto anche per i geni responsabili del catabolismo di altre opine. Ampi settori del plasmidio sembrano ancora silenti, ma ciò può essere dovuto al fatto che controllano caratteri di cui non siamo ancora a conoscenza, tra i quali forse anche la capacità di catabolizzare altre sostanze nutritive.

Le mutazioni a carico del *T*-DNA hanno avuto un particolare interesse per il fatto che hanno permesso di identificare geni trasferiti dal batterio alla pianta. Nel plasmidio per la nopalina studiato da Schell e Van Montagu, una mutazione ha eliminato la sintesi della nopalina nei tumori; essa era localizzata all'interno del *T*-DNA. L'ipotesi preveggente di Morel è stata pertanto confermata: a livello del gene c'era una prova diretta che la capacità di sintetizzare la nopalina fosse controllata dal DNA trasferito dal batterio alla pianta. È risultato che le cellule tumorali stimolate da questo plasmidio *Ti* mutante ospitavano non solo l'abituale *T*-DNA, ma anche una copia del trasposone che aveva dato origine alla mutazione. Ora, il trasposone in questione, il *Tn7*, è un segmento di DNA piuttosto grande. La sua presenza nelle cellule tumorali ha significato che un ampio segmento di DNA estraneo, sufficiente a trasportare diversi geni, può essere trasportato assieme al *T*-DNA. Chiaramente, dunque, doveva essere possibile fare del *T*-DNA un vettore per l'introduzione di un gene desiderabile nelle cellule vegetali. Se un simile gene poteva essere fatto funzionare nelle cellule vegetali e se queste cellule potevano essere rigenerate in piante sane, ecco che si sarebbe avuto a disposizione un metodo di ingegneria genetica vegetale.

Altre mutazioni del trasposone nel *T*-DNA hanno permesso al gruppo di Schilperoort di identificare dei punti che controllano la perturbazione dell'attività ormonale, chiaramente responsabile della crescita tumorale. Le mutazioni determinano due tipi di cambiamenti nella morfologia tumorale: una classe di mutazioni fa emergere delle radici dal tumore, mentre l'altra classe dà origine a germogli. Questo genere di cambiamenti è provocato nel normale tessuto vegetale da uno squilibrio ormonale: un elevato rapporto tra auxine e citochinine induce una eccessiva formazione di radici, mentre un basso rapporto genera un maggior numero di germogli. Schilperoort ha dimostrato che l'effetto delle mutazioni riguardanti la morfologia tumorale poteva in realtà essere capovolto correggendo l'ipotetica deficienza ormonale, cioè aggiungendo citochinine al mezzo in caso di mutanti «con molte radici» e aggiungendo, invece, delle auxine in caso di mutanti «con molti germogli».

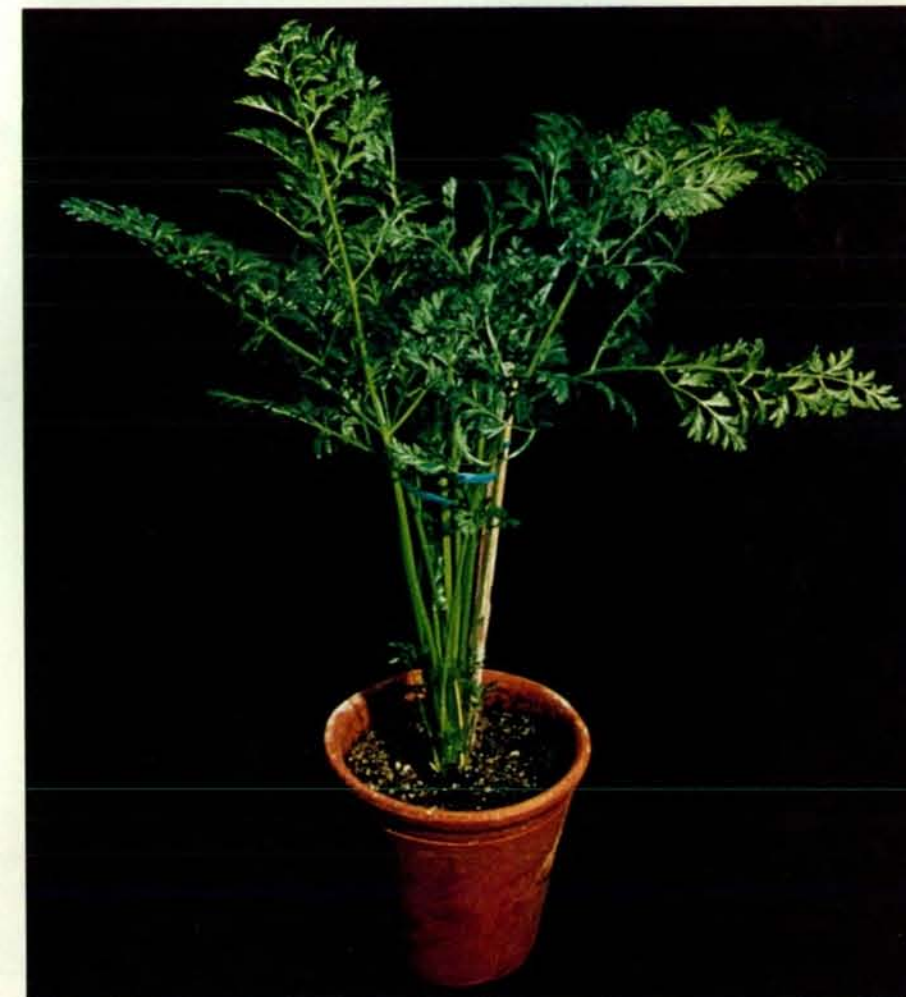
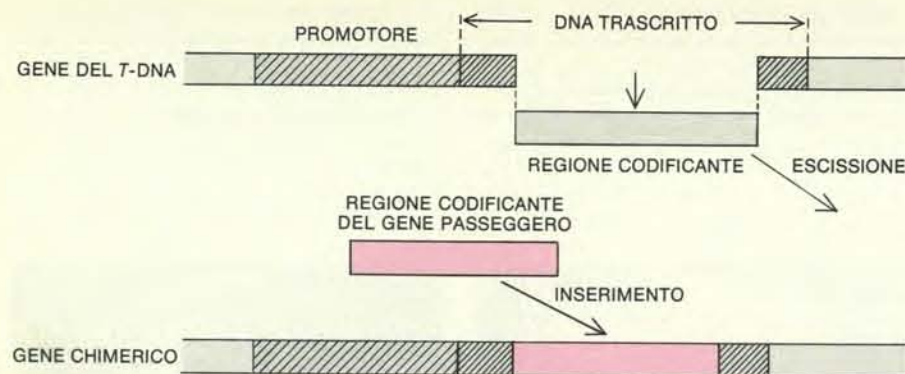
Nel laboratorio di Nester, diversi ricercatori hanno messo a punto una ben particolareggiata mappa del *T*-DNA, includendovi i loci genici sia per il carattere «molte radici» sia per il carattere «molti germogli». Il metodo a cui essi sono ricorsi è stato ancora la mutagenesi del trasposone, ma, dato che ormai si conosceva in una certa misura la struttura del plasmidio *Ti*, tale mutagenesi ha potuto essere «mirata su un determinato sito», cioè è stato possibile scegliere un particolare segmento di *T*-DNA, clonarlo in un piccolo plasmidio vettore, provocare in esso una mutazione e quindi trasferire tale segmento in *A. tumefaciens* e nel plasmidio *Ti*, con una doppia ricombinazione. Si è ottenuto un plasmidio *Ti* con una mutazione in un segmento predeterminato.

Una mutagenesi intensiva del *T*-DNA ottenuta in questo modo ha così permesso di definire i loci per i caratteri «molte radici» e «molti germogli» e inoltre il sito di un'altra classe di mutazioni che generano tumori insolitamente voluminosi. Regioni estese del *T*-DNA sono risultate silenti, cioè le mutazioni verificatesi in esse non hanno avuto effetti evidenti. Fatto curioso, nessuno dei 75 plasmidi mutanti ha perso completamente la sua virulenza, il che implica che tutti i geni che

La mappa genetica di un plasmidio si determina con la mutagenesi del trasposone, un segmento di DNA che tende a saltare a un nuovo sito e che può portare un gene che conferisce la resistenza agli antibiotici. In *Agrobacterium*, mediante coniugazione, viene introdotto un plasmidio «suicida», che contiene un trasposone *(in colore)* per la resistenza alla canamicina *(1)*. Il trasposone può saltare *(2)* su un sito del cromosoma batterico *(a)* o sul plasmidio *Ti (b, c)*. Dovunque esso si integra, provoca una mutazione. Così, procedendo a una selezione per quanto riguarda la resistenza alla canamicina, è possibile isolare cellule in cui il plasmidio *Ti* abbia subito questa mutazione. Ogni cellula viene sottoposta a una prova che ne evidenzi un eventuale cambiamento nella funzione genetica. Il gene per questa funzione può essere determinato confrontando la successione dei segmenti del plasmidio mutante con quella del plasmidio normale *(3)*. Il segmento *x* è scomparso dal plasmidio della cellula *b* e si notano, per contro, nuovi segmenti *y* e *z*, i quali sono il risultato di una scissione provocata dall'enzima *Sma* I in corrispondenza di un nuovo sito nel trasposone. Il gene deve trovarsi nel segmento mancante *x*, la cui identità viene confermata dal fatto che la sua dimensione è uguale alla somma di *y* e *z* meno la dimensione del trasposone.

Si cominciano ora a costruire le mappe genetiche per due plasmidi *Ti*: per l'octopina *(a sinistra)* e per la nopalina *(a destra)*. È stato definito anche il *T*-DNA e, al suo interno, sono stati localizzati i loci genici che controllano la sintesi dell'octopina o della nopalina. Alla stessa stregua sono stati localizzati i geni che influiscono sulla morfologia tumorale e che sono rivelati da mutazioni che provocano abbondanza di radici e di germogli e notevoli dimensioni del tumore.

Un gene estraneo viene inserito in una cellula vegetale e da questa si rigenera una pianta. Un piccolo plasmidio che porta un segmento di *T*-DNA *(in nero)*, il quale include il locus per il carattere «molte radici», viene tagliato e aperto mediante una endonucleasi. In esso viene inserito un segmento di DNA «passeggero»: un gene di lievito per l'enzima alcooldeidrogenasi assieme a un gene per la resistenza alla canamicina (e che serve per la selezione). Il plasmidio ricombinante viene nuovamente clonato in un plasmidio che ha una ampia gamma di ospiti; il nuovo vettore intermedio è introdotto, infine, in una cellula di *A. tumefaciens*, che contiene un plasmidio *Ti* di tipo selvatico (non modificato). Il segmento di *T*-DNA, che è stato così inserito nel vettore intermedio, può ricombinarsi con lo stesso segmento *(in nero)* nel plasmidio *Ti*. Una rara ricombinazione a doppio *crossing over* inserisce il segmento, così manipolato, con il suo passeggero, nel plasmidio *Ti*. Il vettore viene poi espulso dalla cellula introducendo in essa un plasmidio «incompatibile». I batteri che portano nel loro *T*-DNA tale frammento vengono identificati in base alla loro resistenza alla canamicina. Batteri del genere sono successivamente utilizzati per trasformare le cellule delle piante di tabacco. Il *T*-DNA, con il suo passeggero, viene integrato nel DNA della pianta, dando origine a cellule mutanti con abbondanti radici. Queste cellule producono piante sane, le cui cellule trasformate contengono il *T*-DNA nel quale è incluso il gene per l'alcooldeidrogenasi.

controllano la trasmissione del *T*-DNA a una cellula vegetale e la sua inserzione nel DNA cromosomico devono trovarsi al di fuori del *T*-DNA, nella regione responsabile della virulenza, che è posta alla sua sinistra. Sotto questo aspetto, il *T*-DNA è diverso sia dai virus sia dai trasposoni, i quali portano con loro le funzioni vitali per cui possono trasferirsi in un nuovo ospite.

Un gene si esprime quando il suo DNA viene trascritto in un RNA messaggero e l'RNA viene tradotto in proteina. Le cellule del tumore del colletto contengono RNA trascritto a partire dai geni del *T*-DNA. Sembra che tale RNA abbia le caratteristiche speciali dell'RNA messaggero presente nelle cellule eucariote (le cellule nucleate di tutti gli organismi più complessi dei batteri), ma non quelle dell'RNA messaggero delle cellule procariote (cellule batteriche). Sono state determinate le dimensioni e le posizioni sulla mappa di singole molecole dell'RNA tumorale. RNA singoli corrispondono ai loci responsabili dei caratteri «molte radici» e «voluminoso», relativo alla morfologia del tumore, nei plasmidi responsabili della sintesi di octopina e nopalina; due RNA sono codificati invece dal locus responsabile del carattere «molti germogli». Fatto interessante, praticamente tutte le regioni silenti del *T*-DNA codificano anch'esse RNA messaggeri, alcuni dei quali possono corrispondere a geni per le opine che non sono stati ancora scoperti.

Le caratteristiche biologiche uniche del plasmidio *Ti* lo rendono un agente naturale per il trasferimento di geni. Esso realizza l'inserzione del *T*-DNA nel DNA cromosomico della pianta e questo *T*-DNA viene da allora in poi mantenuto dalla pianta ospite, viene duplicato con i geni propri di questa pianta, trascritto in RNA e (almeno in gran parte) espresso sotto forma di proteina. In che modo il plasmidio «decide» quale parte di se stesso è *T*-DNA? La risposta può risiedere in sequenze specifiche che fungono da segnale e che sono state identificate nelle regioni limite ad ambedue le estremità del *T*-DNA. In tali regioni vi è una sequenza di 25 nucleotidi. (Nucleotidi sono le quattro subunità del DNA, la cui successione nelle regioni codificanti di un gene è l'informazione genetica cifrata che specifica una particolare proteina.) Questa sequenza è un segnale di «ripetizione diretta». L'esatta ripetizione di una breve sequenza nucleotidica, sia diretta sia «invertita» serve spesso come segnale per gli eventi di ricombinazione. Le ripetizioni dirette possono definire i limiti del *T*-DNA; qualsiasi DNA estraneo inserito tra le ripetizioni dovrebbe essere trattato come *T*-DNA, cioè dovrebbe essere trasferito nella cellula vegetale.

Descriverò ora una strategia per introdurre un gene estraneo in una pianta mediante un plasmidio *Ti* e *A. tumefaciens*. Si comincia con un plasmidio ricombinante: un piccolo plasmidio adatto per la clonazione nel comune batterio *Escherichia coli*, in cui è stato inserito un segmento conosciuto di *T*-DNA, come nel processo di mutagenesi mirata su un determinato sito, descritta sopra. Il plasmidio ricombinante viene aperto, in corrispondenza di un sito del segmento di *T*-DNA, da un'endonucleasi di restrizione. Un pezzo del DNA «passeggero» viene quindi inserito e saldato nell'apertura. Esso è bipartito: cioè consiste del gene estraneo che si vuole trasferire nella pianta sul *T*-DNA e, attaccato a esso, di un marcatore genetico, cioè di un gene batte-

Le cellule di carota, trasformate da plasmidi *Ri* contenuti in *Agrobacterium rhizogenes*, rigenerano piante intere, sane, di carota. Il *T*-DNA del plasmidio trasforma le cellule di una fetta di carota in radici pelose che crescono rapidamente *(in alto a sinistra)*. Una singola radice messa in un terreno colturale prolifera *(in alto a destra)*. Quando la coltura viene fatta su terreno solido per ottenere un callo e questo è trasferito in un terreno liquido, si forma un embrione che si sviluppa in pianta *(in basso)*. Dalle radici pelose si rigenerano anche altri tipi di pianta. Sembra che il plasmidio *Ri* sia un vettore di geni reso naturalmente innocuo, che può dare origine a piante sane modificate geneticamente. Queste fotografie sono di Jacques Tempé dell'Università di Parigi a Orsay.

Per far sì che un gene passeggero venga espresso dalle cellule vegetali è necessario costruire un gene chimerico. La regione codificante di un gene del *T*-DNA viene isolata, lasciando indisturbato il promotore che precede il gene (e regola la sua trascrizione nell'RNA) e le sequenze segnale *(tratteggio più fitto)* che lo fiancheggiano (e possono anche regolare la trasformazione dell'RNA per la sua traduzione in proteina). La regione codificante del gene passeggero viene quindi inserita tra le regioni regolatrici, cosicché si esprimerà come fosse parte integrante del *T*-DNA.

rico per la resistenza alla canamicina. Questo plasmidio combinato viene quindi nuovamente clonato in un plasmidio più grande, che è introdotto in un ceppo di *Agrobacterium* contenente un plasmidio *Ti* di tipo selvatico (cioè non modificato). Con il proliferare di questo genere di batteri, talvolta si verifica un raro evento di doppia ricombinazione, il che porta a un clone di batteri il cui *T*-DNA contiene un inserto: il DNA «passeggero» introdottovi mediante manipolazione genetica. Questi batteri possono essere identificati e selezionati grazie alla loro sopravvivenza in un mezzo contenente canamicina, dopo che è stato espulso il plasmidio vettore ricombinante. I batteri selezionati sono quindi introdotti in una pianta e il loro *T*-DNA viene inserito nelle cellule di quest'ultima in associazione con il DNA «passeggero».

Numerosi esperimenti hanno chiarito che questa strategia, o una delle molte strategie analoghe, può dare origine a cellule di tumore del colletto geneticamente modificate, che crescono su una piastra di coltura. Questo risultato è ancora ben lungi dall'obiettivo di riuscire a far crescere piante sane, geneticamente modificate, partendo da una simile linea cellulare trasformata. Poche piante complete sono state ottenute partendo da cellule non modificate di tumore del colletto e, in ogni caso, esse sono risultate mutanti il cui *T*-DNA era stato in tutto o in parte eliminato. La conclusione è che la presenza di geni del *T*-DNA è incompatibile con lo sviluppo di piante sane, presumibilmente perché essi portano delle perturbazioni nelle concentrazioni delle auxine e delle citochinine. In che modo allora si può rigenerare una pianta sana con il *T*-DNA e il suo passeggero integrati nel DNA del cromosoma della pianta?

Un possibile modo è stato suggerito da alcuni sorprendenti risultati ottenuti di recente da Andrew N. Binns dell'Università della Pennsylvania e da Kenneth A. Barton, Antonius J. M. Matzke e me nel mio laboratorio presso la Washington University di St. Louis. Abbiamo inserito un gene di lievito, che codifica per l'enzima alcooldeidrogenasi, nel locus del plasmidio per la nopalina che controlla il carattere «molte radici» e, con il risultante plasmidio, abbiamo infettato una pianta di tabacco. Si è così ottenuta una coltura di cellule mutanti con molte radici. Si è avuta pòi la sorpresa di scoprire che tali cellule non erano tipicamente tumorali: per crescere abbísognavano dell'aggiunta di citochinina. Inoltre, sintetizzando nopalina e non avendo bisogno di alcuna aggiunta di auxina, esse hanno dato origine a piante di tabacco complete. Il DNA isolato da tali piante conteneva una ventina di copie di *T*-DNA per cellula, ciascuna delle quali includeva una copia del gene eucariota per l'alcooldeidrogenasi, che era stato inserito. In altre parole, le nostre piante, ben lungi dall'essere mutanti per delezione, avevano un primato assoluto di numero di copie di *T*-DNA per cellula. Dopo l'autoimpollinazione, esse hanno prodotto semi che si sono sviluppati in piante sane, le cui cellule contenevano *T*-DNA completo, in cui era incluso ancora, in molteplici copie, il gene del lievito.

Concludiamo così che una mutazione che determina la comparsa di molte radici in qualche modo «disarma» il *T*-DNA per la nopalina per cui le cellule di tabacco, trasformate dal *T*-DNA possono rigenerarsi. Rimane da vedere se questo discorso vale anche per altri tipi di plasmidi *Ti* e in altri tipi di piante.

Un modo più generale per procurarsi un vettore *T*-DNA «disarmato» potrebbe essere quello di eliminare tutti i geni propri del *T*-DNA, lasciando solo il DNA «passeggero» e le sequenze non codificanti, che fungono da segnali per la ripetizione diretta. Tali vettori sono stati di recente «costruiti» in vari laboratori, incluso il mio. Essi possono risultare i vettori di scelta per la maggior parte delle piante suscettibili di essere infettate con *A. tumefaciens*.

Esiste un'altra famiglia di plasmidi di *Agrobacterium* che possono anch'essi servire da vettori di geni e che conducono alla formazione di piante sane geneticamente modificate. Questi plasmidi provocano una proliferazione delle radici, chiamata malattia delle radici pelose, in piante infettate da *A. rhizogenes* e sono detti plasmidi *Ri («root inducing»)*. Le radici, come il tessuto del tumore del colletto, si sviluppano rapidamente in coltura in assenza del batterio infettante. Jacques Tempé e collaboratori dell'Institut National de la Recherche Agronomique hanno dimostrato che le cellule radicali contengono le opine e parecchie copie di un *T*-DNA. La cosa importante è che i plasmidi *Ri*, come i plasmidi *Ti*, che presentano il carattere mutante «molte radici», danno origine a cellule vegetali trasformate, che possono rigenerarsi in piante sane, integre, fertili. Sembra che i plasmidi *Ri* mostrano di essere vettori naturalmente disarmati.

Dato che geni estranei possono venir introdotti in certe piante e che possono venir rigenerate piante sane geneticamente modificate, il successivo interrogativo è se quei geni saranno espressi, cioè daranno origine a RNA messaggero e proteina, e avranno così un effetto nella loro nuova sistemazione. Numerosi esperimenti in cui sono stati trasferiti geni batterici, di lieviti, animali e vegetali non hanno avuto successo sotto questo aspetto. Il nostro gene di lievito responsabile della sintesi dell'alcooldeidrogenasi, per esempio, non si esprime nelle piante di tabacco rigenerate (cioè l'enzima non viene sintetizzato). L'espressione di un gene è regolata da varie sequenze che fungono da segnali e che fiancheggiano il DNA che codifica per una proteina. Le sequenze regolatrici più importanti sono localizzate nella regione del promotore, a monte del gene. I segnali che circondano un gene estraneo inserito devono essere riconoscibili dalla cellula vegetale, se deve aver luogo l'espressione di quel gene. Da diversi anni è possibile far sintetizzare ai batteri proteine animali, inserendo in essi un gene «chimerico», cioè un gene in cui l'informazione riguardo alla regolazione, prelevata da un gene batterico, viene accuratamente inserita e saldata nella regione codificante di un gene animale.

In epoca molto recente, Michael W. Bevan nel mio laboratorio, il gruppo di Gand in Belgio e ricercatori della Monsanto Company di St. Louis sono riusciti a produrre un analogo gene chimerico che si esprime nelle cellule vegetali. In tutti e tre i laboratori il procedimento è costituito nell'asportare, da un gene del *T*-DNA, mediante escissione, la regione codificante, nell'inserire (saldandolo) un gene batterico per la resistenza alla canamicina, tra le regioni regolatrici di quel gene, e quindi nell'introdurre questo nuovo gene chimerico nei plasmidi *Ti*, più o meno nel modo che è stato descritto sopra. Le cellule vegetali trasformate dai plasmidi hanno espresso il gene batterico: hanno dimostrato di essere resistenti a una concentrazione di canamicina che avrebbe ucciso le cellule in coltura non protette. Questo successo indica che il problema dell'espressione genica in una pianta modificata geneticamente può, almeno in teoria, essere risolto. A mano a mano che aumenta la conoscenza sui meccanismi di regolazione dei geni nelle piante, dovrebbe essere possibile non solo arrivare alla loro espressione, ma anche raggiungere un'espressione selettiva, modificando la regione del promotore di questi geni. Si potrebbe fare in modo che un gene si esprima solo nelle foglie o solo nei semi, o solo in risposta a un segnale chimico.

Finora non è stato possibile realizzare nello stesso esperimento sia la rigenerazione sia l'espressione dei geni, ma ciò sarà sicuramente possibile tra breve. Il problema sarà allora di riuscire a scegliere e a isolare geni che hanno la probabilità di migliorare le piante coltivate. Vi sono in proposito degli obiettivi ovvii, come le maggiori rese, la resistenza agli agenti infestanti e la tolleranza alle basse temperature e alla siccità. Il guaio è che non c'è alcun modo conosciuto per trovare dei geni che controllino questi caratteri, la maggior parte dei quali è probabilmente controllata da numerosi geni e forse da elaborati meccanismi di regolazione. Ma anche se si potessero identificare i geni appropriati, non c'è alcuna tecnica nota per isolarli mediante i metodi del DNA ricombinante; finora sono stati isolati solo geni i cui prodotti proteici sono noti.

D'altro canto, un obiettivo realizzabile nel prossimo futuro è l'introduzione nelle piante coltivate di nuovi geni estranei che conferiscono resistenza agli erbicidi o ai pesticidi. Se una pianta potesse essere coltivata in presenza di potenti sostanze chimiche per uso agricolo dovrebbe dare una resa più elevata per la minor competizione da parte delle erbe infestanti o per la minore predazione da parte di animali.

La promessa dei plasmidi *Ti* e *Ri* non va oltre la modificazione delle dicotiledoni, che l'*Agrobacterium* infetta. Questo batterio non infetta, invece, le monocotiledoni, tra cui le graminacee e i cereali in particolare. E proprio monocotiledoni sono la maggior parte delle più importanti colture alimentari mondiali. Si continua a non capire la ragione della loro immunità al batterio: si hanno ancora scarse informazioni sul modo in cui il *T*-DNA migra dal plasmidio al nucleo della cellula vegetale e sul modo in cui (o addirittura se) il *T*-DNA sceglie i punti in cui si inserisce nel DNA della pianta. In questa importante area della ricerca è necessario compiere ancora molto lavoro, perché una migliore comprensione della sequenza degli eventi nell'infezione può fare intuire quale fase manca nelle monocotiledoni. Molte specie di dicotiledoni rigenerano facilmente piante intere a partire da cellule coltivate in vitro; tra le monocotiledoni la rigenerazione è invece più rara e più difficile da realizzare, ma di recente sono stati compiuti progressi anche in questo campo nel caso di importanti piante coltivate a partire da singole cellule. Ciò dovrebbe spronare a tentare di allargare la gamma degli ospiti del plasmidio *Ti* anche alle monocotiledoni e a mettere a punto nuovi vettori che siano in grado di funzionare in queste piante.

L'*Agrobacterium* ha mostrato che è possibile inserire DNA estraneo nei cromosomi delle piante. Sono in grado i biologi di imitare il processo naturale con un plasmidio di loro scelta o da loro stessi progettato, oppure i vantaggi della combinazione simbiotica di *A. tumefaciens* e del plasmidio *Ti*, che l'evoluzione e la selezione naturale hanno modellato, sono inimitabili? Finora i finti plasmidi introdotti nelle cellule vegetali con una varietà di tecniche non sono riusciti a integrare il loro DNA nel DNA ospite. Perfino lo stesso plasmidio *Ti* quasi fallisce quando viene artificialmente introdotto nei protoplasti di tabacco (cellule denudate, cioè private della rigida parete esterna). L'efficienza dell'inserimento del *T*-DNA è molto bassa e le estremità del segmento inserito sono anomale, facendo pensare che non funzioni il normale meccanismo di integrazione.

È sicuro, invece, che *Agrobacterium* sa come inserire il proprio *T*-DNA nel nucleo della cellula vegetale con un'efficienza elevata. Come? Può esservi un enzima batterico che asporta mediante escissione il *T*-DNA dal plasmidio. La forma che questo *T*-DNA presenta quando arriva al nucleo deve essere importante: potrebbe essere lineare o circolare, a filamento singolo o a duplice filamento, nudo o complessato con proteine batteriche; potrebbe replicarsi indipendentemente nella cellula vegetale oppure potrebbe integrarsi immediatamente. La cellula batterica infettante potrebbe porre in una certa misura la cellula vegetale in uno stato recettivo che le permetta di ricevere nel proprio DNA l'inserimento del *T*-DNA. Oppure il batterio potrebbe avere una speciale affinità per le cellule recettive. Le sequenze che fungono da segnale e che sono situate alle estremità del *T*-DNA possono essere importanti per l'escissione dal plasmidio o per l'inserimento del DNA della pianta; non è noto però quale struttura (e in quale organismo) legga i segnali.

Per tutti i suoi misteri, il plasmidio *Ti* è oggi l'unico vettore per geni che possa essere utilizzato con successo nelle piante superiori. Esso è diventato un importante strumento di indagine, a parte la vantaggiosa prospettiva del suo sfruttamento per il miglioramento delle piante coltivate. Fornisce un mezzo per manipolare i geni delle piante e pertanto per studiarne la struttura e la regolazione. Ha rivelato l'esistenza di geni che influenzano i livelli ormonali nelle piante e tali geni saranno importanti nello studio dello sviluppo. Infine, il *T*-DNA ha dimostrato per la prima volta che il DNA del plasmidio batterico si ricombina con quello dei cromosomi della pianta, rivelando così, per la trasmissione dei geni, un canale evolutivo in precedenza sconosciuto.

Caldere vulcaniche giganti

Sono crateri con diametri di diverse decine di chilometri formati da eruzioni vulcaniche di gran lunga più violente e catastrofiche di qualsiasi eruzione avvenuta in epoca storica

di Peter Francis

L'eruzione del 18 maggio 1980 del Mount St. Helens, posto nella parte meridionale dello stato di Washington, portò all'emissione di seicento milioni di metri cubi di magma e formò un cratere del diametro di due chilometri. Se tale eruzione viene giustamente considerata un evento altamente spettacolare, che cosa dire, allora, dell'eruzione che avvenne circa 600 000 anni fa 950 chilometri a est del Mount St. Helens stesso? Essa, infatti, produsse 1000 miliardi di metri cubi di pomici e ceneri e formò una caldera, ossia un cratere vulcanico di notevoli dimensioni, di forma allungata, che misurava nella dimensione maggiore 70 chilometri. Nascosti dalla vegetazione e alterati dalla glaciazione, i segni dell'eruzione non sono oggi facilmente individuabili. Il resto più evidente è senz'altro l'Old Faithful geyser dello Yellowstone National Park. Yellowstone è il risultato di processi vulcanici operanti a grandissima scala: è, infatti, una «caldera risorgente» ossia quel tipo di caldera il cui fondo, nei millenni successivi all'eruzione, si incurva lentamente all'insù, dando luogo alla formazione di un duomo. Le caldere risorgenti sono di gran lunga le più grandi che si trovano sulla Terra e ciò porta indubbiamente a pensare che eruzioni vulcaniche simili a quella che formò la caldera di Yellowstone siano tra le più immani catastrofi naturali che possano accadere, paragonabili, forse, solo all'impatto con un asteroide.

Per buona sorte sono rare, tanto è vero che, nelle poche migliaia di anni dell'epoca storica, non se n'è formata nessuna; negli Stati Uniti ne sono state identificate solo tre formatesi nell'ultimo milione di anni. Oltre a quella di Yellowstone, un'eruzione di 700 000 anni fa formò in California la caldera di Long Valley e un'altra, di un milione di anni or sono, la caldera di Valles nel New Mexico. (La caldera di Long Valley è stata studiata recentemente con grande interesse a causa di una serie di piccoli terremoti registrati nell'area.) Come negli Stati Uniti, anche nel resto del mondo si cerca di identificare eruzioni di questo tipo, ma sembra probabile che, nell'ultimo milione di anni, non se ne siano verificate più di dieci in tutto il mondo. D'altra parte, un rilevamento cartografico dettagliato condotto da Thomas A. Steven e da Peter W. Lipman dell'US Geological Survey riporta almeno 18 caldere di età compresa tra i 20 e i 30 milioni di anni nella zona delle San Juan Mountains nel Colorado e numerose altre di età comparabile nel New Mexico, nell'Arizona e nel Nevada. Negli ultimi dieci anni, i vulcanologi hanno compiuto rapidi progressi nella comprensione delle origini delle gigantesche caldere risorgenti e delle catastrofiche eruzioni che le formano.

Il meccanismo fondamentale alla base della formazione delle caldere è oramai sufficientemente chiaro. L'improvvisa eiezione di grandi quantità di magma da una camera magmatica, che si trovi ad alcune migliaia di metri di profondità, provoca lo sprofondamento del tetto lasciando una caldera in superficie. In un classico lavoro sulla geologia dei vulcani in Scozia, pubblicato nel 1909, C. T. Clough, H. B. Maufe ed E. B. Bailey, descrivendo per la prima volta la formazione di una caldera, la definirono «caldera di collasso o di sprofondamento». Questo processo ha luogo a scale diverse e quindi il diametro delle caldere può variare da qualche chilometro a 50 e anche più.

A prescindere, comunque, dalla loro estensione, caratteristica fondamentale di una caldera risorgente è il lento sollevamento del suo fondo causato probabilmente dall'intrusione successiva di magma nella camera magmatica originaria. L'entità di questo sollevamento verticale può superare anche i 1000 metri. A differenza di un vulcano comune, una caldera risorgente è quindi una grande depressione con un sollevamento centrale. Il fenomeno della risorgenza venne individuato per la prima volta nel 1939 dal geologo olandese R. W. van Bemmelen mentre studiava la caldera di Toba, nella parte settentrionale dell'isola di Sumatra. Van Bemmelen pensò che il fondo della caldera fosse stato interessato da un fenomeno di subsidenza con un abbassamento di circa 2000 metri e la conseguente formazione di un lago che successivamente si sarebbe sollevato di diverse centinaia di metri con la formazione dell'isola di Samosir al centro del lago. La caldera risorgente di Toba resta a tutt'oggi la più grande tra quelle note; nella dimensione maggiore si estende per quasi 100 chilometri. Ciononostante non è stata studiata a fondo e le maggiori conoscenze sul fenomeno della risorgenza (ricordo che il termine «risorgenza» fu introdotto nel 1962 da Robert L. Smith e R. S. Bailey) si devono in gran parte ai geologi del Geological Survey che hanno studiato le caldere degli Stati Uniti sudoccidentali.

Un'ulteriore caratteristica delle caldere risorgenti può essere ricondotta alla natura dei processi vulcanici. In ogni eruzione vulcanica, il magma che raggiunge la superficie può essere emesso in tre maniere: sotto forma di lava, di materiale piroclastico lanciato nell'atmosfera e che poi ricade (come quello che, nel 79 d.C.

La caldera di Cerro Galán, nell'Argentina nordoccidentale, non era mai stata rilevata dalle comuni ricognizioni geologiche dell'area, ma venne scoperta in questa fotografia a falsi colori ripresa da un satellite Landsat alla quota di 900 chilometri. Le dimensioni stesse della caldera rendevano difficile la sua individuazione. Il bordo della caldera, in parte ricoperto da neve, ha forma ellittica e l'asse maggiore lungo la direzione nord-sud misura 34 chilometri. Il centro della caldera, anch'esso innevato, è un rilievo la cui cima è alta 6000 metri sul livello del mare (un centro di risorgenza). Il rilievo è dovuto all'accumularsi di magma nei millenni successivi all'eruzione catastrofica che ha spinto verso l'alto il fondo della caldera. Il terreno intorno alla parte settentrionale della caldera, caratterizzato da profonde scanalature, rappresenta i resti molto erosi di una enorme estensione di ignimbrite, una roccia che si è formata da pomice e cenere emesse nel corso dell'eruzione.

ricoprì Pompei) e, infine, di nube ardente (colata piroclastica). La lava è, sostanzialmente, magma che raggiunge la superficie, si sparge come un liquido e solidifica in rocce ignee che possono essere a struttura cristallina fine o addirittura vetrosa. A seconda della composizione del magma, le rocce vanno dai basalti, grigio scuri e con basso contenuto in silice (SiO_2), alle andesiti, alle daciti fino alle rioliti grigio chiare, particolarmente ricche in silice. Il materiale piroclastico lanciato nell'atmosfera consiste, invece, di pomici (piroclastiti leggere e vetrose) insieme a particelle più piccole di cenere e polvere. (Il termine «polvere» viene generalmente usato per le particelle che misurano meno di quattro micrometri di diametro, mentre quelle comprese tra quattro e 63 micrometri sono dette «ceneri fini».) Questo materiale si forma quando frammenti di magma solidificato sono trasportati in alto nell'atmosfera da correnti convettive di gas molto caldi.

Anche le nubi ardenti sono costituite da pomici, ceneri e polveri, ma in questo caso il materiale vulcanico lambisce il suolo e, sostenuto e fluidificato da gas a elevatissima temperatura, si muove alla velocità di oltre 100 metri al secondo. Spesso la nube deposita i clasti, o frammenti, di pomice a diverse decine di chilometri di distanza dalla sorgente dell'eruzione.

La maggior parte delle eruzioni vulcaniche produce lava e ceneri lanciate nell'atmosfera, mentre le nubi ardenti sono meno comuni. Eppure nelle eruzioni che formano le caldere risorgenti costituiscono certamente la parte preponderante e il loro materiale si accumula spesso nella caldera appena formata raggiungendo uno spessore che può superare anche i 1000 metri. Alla base di queste pile, i clasti di pomice si ammorbidiscono e si saldano insieme dando origine a una inconfondibile formazione rocciosa. Nel 1935 P. Marshall, studiando le rocce prodotte da nubi ardenti, in Nuova Zelanda, le chiamò «ignimbriti» nome che significa appunto rocce formate da nubi ardenti. Il termine si usa oggi per tutti i depositi pomicei dovuti a nubi ardenti, siano i clasti saldati o no. D'altra parte, grossi spessori di ignimbriti sono uno dei migliori indizi per localizzare antiche caldere.

Le caldere risorgenti sono, dunque, molto estese e il materiale emesso durante le eruzioni è estremamente voluminoso; ciononostante, non è molto facile identificarle. In molti casi sono, per così dire, camuffate dalla loro stessa enorme estensione. Pochi anni fa, M. C. W. Baker e io, lavorando inizialmente in Inghilterra presso la Open University e successivamente al Lunar and Planetary Institute di Houston, conducemmo una ricerca sulle caldere della Cordigliera delle Ande, in America Meridionale, analizzando le immagini riprese dai satelliti Landsat. La ricerca poteva usufruire dell'enorme vantaggio che l'immagine ripresa dal satellite offre sulle altre effettuate, per esempio, dall'aereo: l'altezza del satellite, infatti - mediamente sui 900 chilometri - fornisce una visione sinottica delle grandi strutture geologiche della Terra, altrimenti non ottenibile. In tal modo, siamo riusciti a identificare due enormi caldere che non erano mai state rilevate da altre indagini geologiche.

La più impressionante delle due è la caldera di Cerro Galán, in Argentina, che ha un diametro di 34 chilometri ed è circondata da una spettacolare estensione di ignimbrite che si protende per ben 70 chilometri dall'orlo della caldera stessa. Abbiamo comunque scoperto più tardi che la caldera era già stata notata da J. D. Friedman del Geological Survey e da Grant H. Heiken del Johnson Space Center nelle fotografie riprese dallo *Skylab 4* nel 1974. La caldera è così giovane (2,6 milioni di anni) e la sua struttura così poco ambigua che noi abbiamo dubitato ben poco della sua natura.

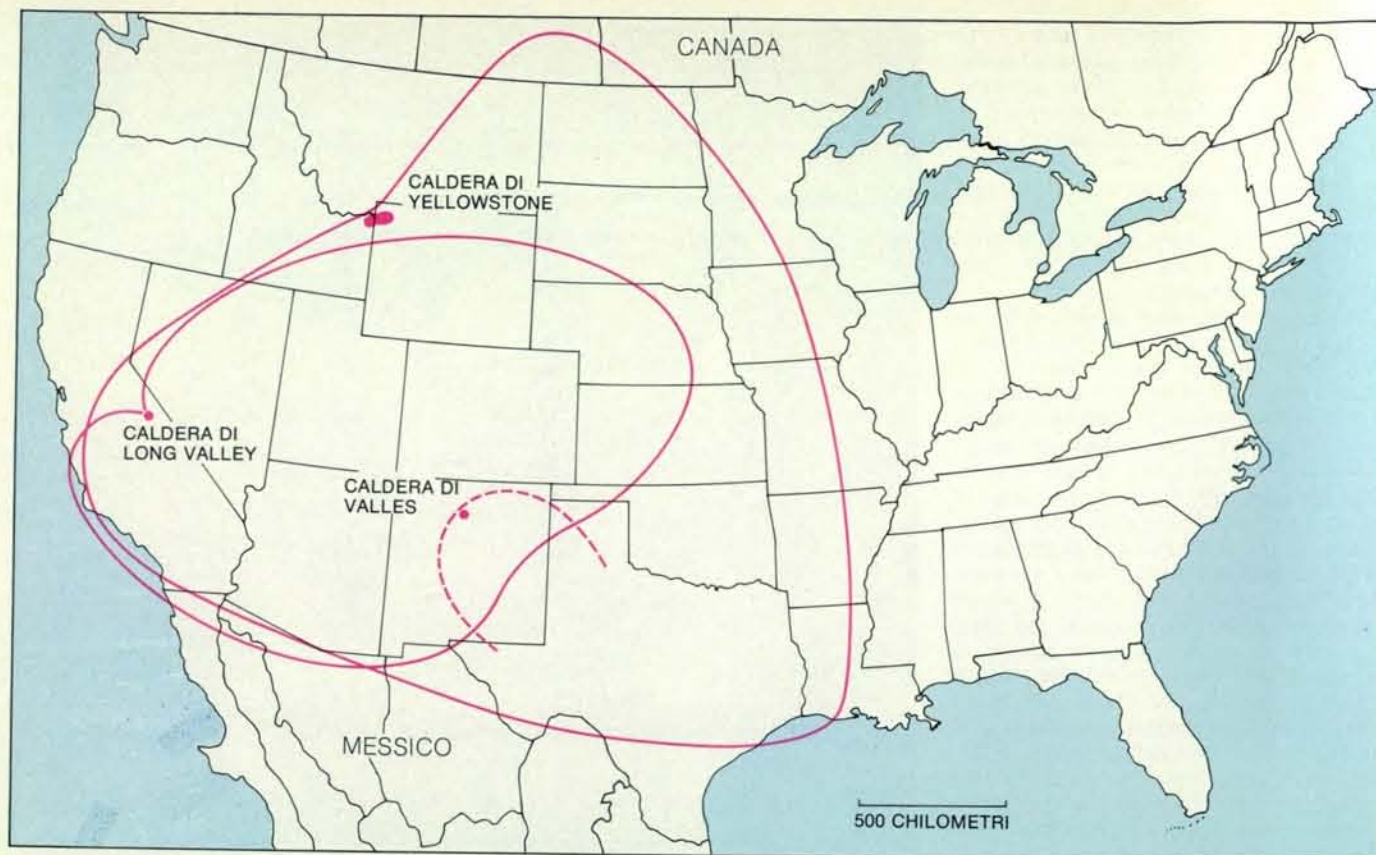
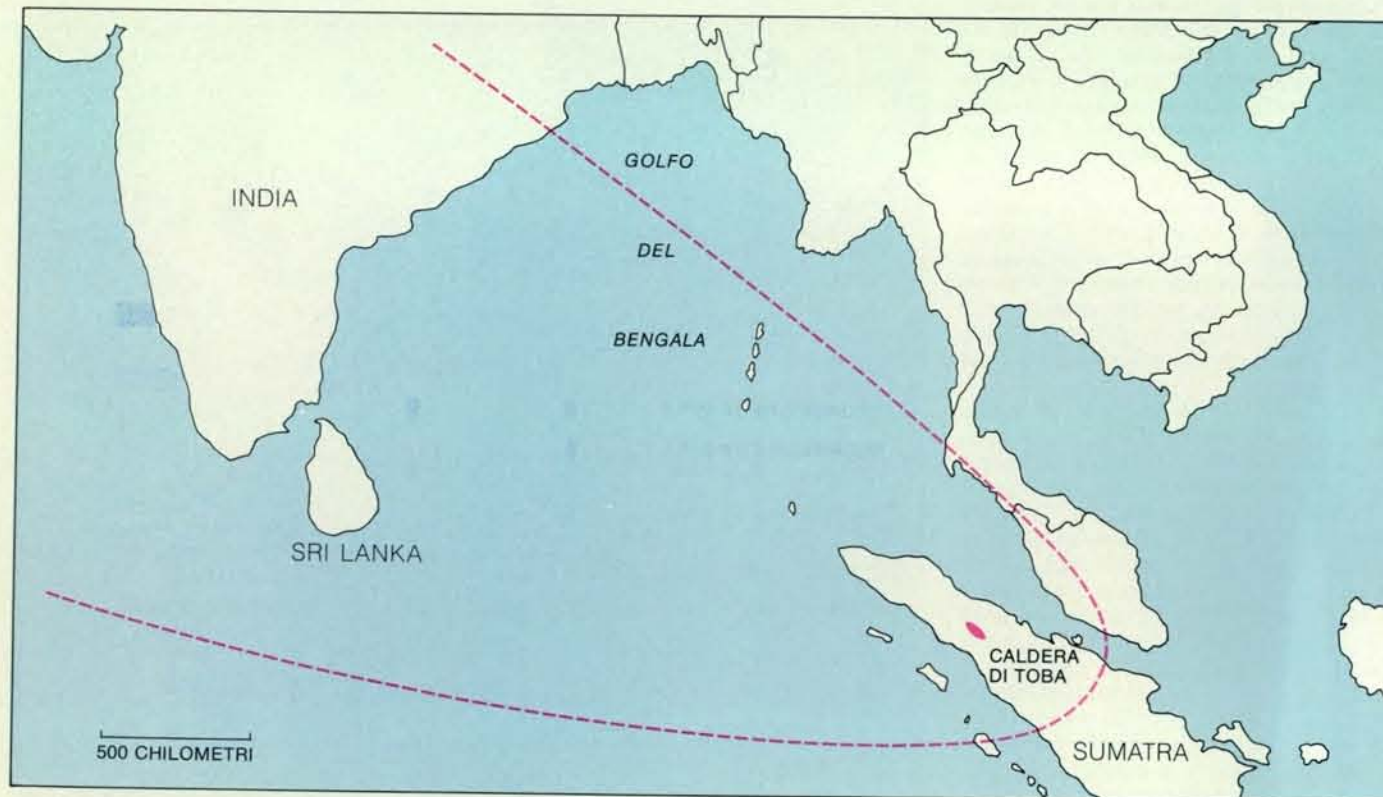
Ciò non si può dire, invece, della seconda caldera. Le immagini Landsat sembrano indicare che il massiccio del Kari Kari, in Bolivia, alto 5000 metri, rappresenti con tutta probabilità il centro di risorgenza di una vecchia ed estesa caldera. Per la verità, mappe già disponibili del Kari Kari facevano pensare che si trattasse di un batolite: una grande massa di roccia ignea fatta di grossi cristalli, solidificata nella crosta e successivamente esposta all'erosione. Il lavoro che conducemmo direttamente sul posto confermò la nostra identificazione: la struttura della roccia, infatti, ben saldata, rivela inconfondibilmente la sua natura ignimbritica. Il centro di risorgenza è ciò che resta di un enorme cratere che doveva essere largo ben 36 chilometri nella sua dimensione massima e la cui età dovrebbe essere di circa 20 milioni di anni. Oltre a queste caldere, se ne sono trovate anche altre nelle Ande centrali, ma poiché sono localizzate in regioni tuttora inesplorate sarà molto lungo e difficile studiarle in loco.

Inizialmente, si potrebbe pensare che la regione più adatta per un'eruzione che formi una caldera risorgente sia una zona di subduzione che, come è noto, rappresenta il margine dove una zolla di crosta oceanica scivola sotto una zolla di crosta continentale e si immerge nel sottostante mantello. In effetti le zone di subduzione sono sedi di intense attività sismiche e vulcaniche, ma mentre la caldera di Toba, nell'isola di Sumatra, è in una siffatta regione, nella maggior parte dei casi ne siamo piuttosto lontani. Quasi tutte le più giovani caldere degli Stati Uniti, ad esempio, distano centinaia di chilometri da una qualsiasi zona nota di subduzione.

Ciò malgrado, le caldere risorgenti non sono affatto distribuite a caso sulla superficie terrestre. Difatti, le ignimbriti che le caratterizzano sono prodotte da eruzioni di magmi dacitici o riolitici, viscosi, ricchi in silice ed emessi generalmente in aree dove la crosta continentale è spessa. Queste considerazioni portano, allora, a pensare che le caldere risorgenti tendano a formarsi in regioni della crosta continentale dove i cosiddetti «punti caldi» *(hot spot)* siano abbastanza vasti e duraturi da fondere considerevoli volumi di roccia. In realtà, il punto caldo non fonde direttamente la crosta continentale, ma solo una parte del mantello creando in questo modo copiosi magmi basaltici che, risalendo verso la superficie terrestre, sono in grado di fondere anche le rocce localizzate a livelli superiori.

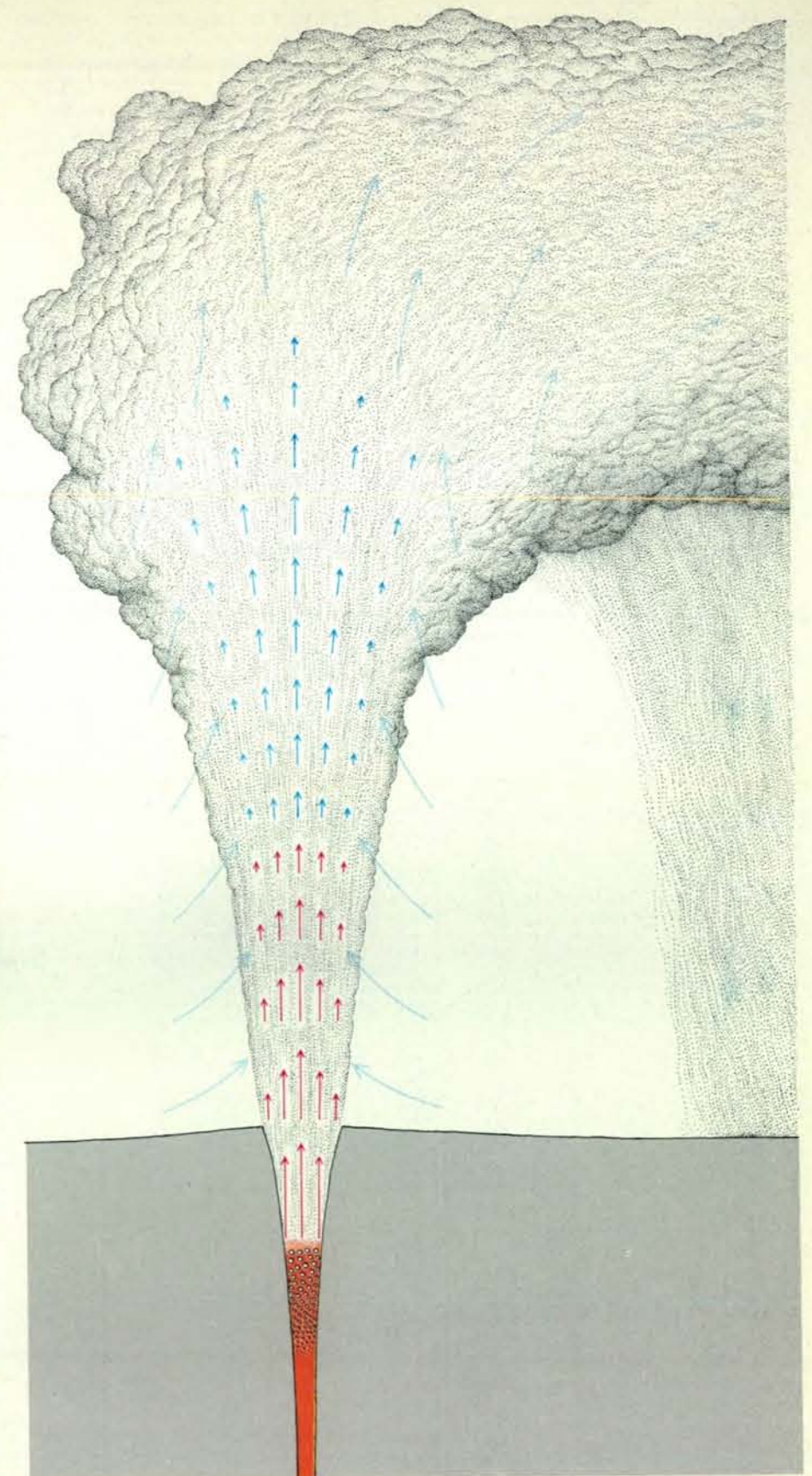
Questa scarpata di ignimbrite di 500 metri è stata messa a nudo dall'erosione a Cerro Galán, circa 20 chilometri a ovest dell'orlo della caldera. Ogni strato orizzontale rappresenta il materiale depositato da una eruzione. Lo strato più elevato venne deposto dall'eruzione che formò la caldera.

Questa ignimbrite «saldata», raffigurata in dimensioni reali, è stata fotografata sui fianchi della caldera di Valles, nel New Mexico. Le macchie scure sono clasti, ossia frammenti, di pomice vetrosa (roccia vulcanica spugnosa) compressi e fusi dal calore del materiale emesso dall'eruzione che ha formato la caldera. I clasti si trovano ora incorporati in una roccia rigida e compatta.

Questa ignimbrite non saldata è stata fotografata sui fianchi della caldera di Cerro Galán. I clasti di pomice giacciono in una matrice di polvere e cenere vulcaniche e sono rimovibili con le mani.

In questa cartina degli Stati Uniti sono riportate le zone interessate dalla caduta di cenere dovuta alle tre eruzioni vulcaniche che hanno formato caldere risorgenti nell'ultimo milione di anni: quella di Valles, nel New Mexico (1 000 000 di anni fa); quella di Long Valley, in California (700 000 anni fa) e quella di Yellowstone, nello Wyoming (600 000 anni fa). Questo lavoro cartografico è stato compiuto in gran parte da G. A. Izett e colleghi dell'US Geological Survey; per le ceneri della caldera di Valles la linea è tratteggiata poiché non è certa la delimitazione dell'area. I vulcani comuni possono spargere ceneri su vaste aree per spessori di frazioni di millimetro, mentre le eruzioni connesse alla formazione delle caldere danno origine a strati di cenere che possono presentare spessori dell'ordine dei centimetri.

La cenere della caldera di Toba nell'isola di Sumatra cadde 75 000 anni fa su gran parte del Golfo del Bengala e su ampie zone di Sumatra, di Sri Lanka e dell'India. L'andamento della sua distribuzione è stata determinato da un gruppo diretto da Dragoslav Ninkovich del Lamont-Doherty Geological Observatory della Columbia University. È stato trovato uno strato di cenere spesso oltre 10 centimetri in campioni estratti dal fondo marino a più di 2000 chilometri di distanza dalla caldera. Quella di Toba è la più grande e la più recente caldera conosciuta.

Consideriamo adesso alcune caldere in particolare. Negli Stati Uniti la caldera di Yellowstone chiude a nord-est una serie di zone di attività vulcanica che inizia nell'Idaho con le rocce basaltiche della piana dello Snake River. Durante gli ultimi 15 milioni di anni, il centro dell'attività vulcanica si è spostato lungo questa traccia fino a raggiungere lo Wyoming, sua attuale posizione, in risposta probabilmente al movimento della zolla che comprende la crosta continentale dell'America Settentrionale, al di sopra di un punto caldo stazionario del mantello. Molte altre caldere, tutte di non più di qualche decina di milioni di anni, si trovano in un'area di parecchie centinaia di chilometri quadrati tra il Nevada, l'Arizona, lo Utah e il New Mexico. Di queste, la più recente segue i fianchi del rift del Rio Grande che corre verso nord per centinaia di chilometri attraverso il New Mexico fino al Colorado. Il rift si sarebbe prodotto per assottigliamento della crosta continentale e i rift presenti nella crosta oceanica intorno a molti sistemi di archi insulari del Pacifico dovrebbero aver avuto analoghi processi di formazione. In quello del Rio Grande, il forte assottigliamento potrebbe aver fatto sì che il calore proveniente dal mantello agisse con grande effetto.

Altri meccanismi sono, però, possibili. In Argentina e in Bolivia si sono formate caldere risorgenti non solo lungo la principale cordigliera vulcanica andina, ma anche in una seconda cordigliera spostata di oltre 200 chilometri verso l'interno. In questa regione non si trova alcun segno di assottigliamento crostale; anzi, lo spessore della crosta continentale al di sotto delle caldere di Cerro Galán e di Kari Kari è compreso tra i 40 e i 50 chilometri. Si suppone, quindi, che in tal caso i condotti del magma si siano sviluppati sotto l'azione di sforzi distensivi locali: sarebbero, cioè, il risultato di una fratturazione della crosta causata dalla pressione del magma in ascesa.

L'evoluzione di una caldera è, generalmente, caratterizzata da una successione di eventi, quali la formazione di un duomo precalderico, il collasso della caldera, eruzioni di materiale piroclastico lanciato nell'atmosfera e nubi ardenti, una risorgenza postcalderica e, infine, tarde estrusioni laviche. Per duomo precalderico si intende quel sollevamento della superficie terrestre che precede una massiccia eruzione e che avviene quando un grande volume di magma si intrude a un livello piuttosto superficiale della crosta continentale creando un plutone, o camera magmatica, la cui sommità è posta, in genere, a soli 4000 o 5000 metri al di sotto della superficie terrestre. Tale processo genera un campo locale di sforzi distensivi che agisce sulle rocce più superficiali e che risulta di fondamentale importanza per l'evento successivo: il collasso della caldera.

In una eruzione vulcanica si produce una colonna pliniana quando la velocità dell'eruzione è elevata, quando il magma eruttato contiene grandi quantità di gas disciolti e quando la bocca eruttiva è stretta, per esempio, dai 50 ai 100 metri di diametro. Alla profondità di circa 1000 metri il gas sfugge dalla soluzione, causando una frammentazione esplosiva del magma in pomici e ceneri incandescenti. Le particelle sono lanciate in alto balisticamente *(frecce in rosso)* e formano la parte più bassa della colonna pliniana. Correnti convettive le trasportano poi più in alto di parecchi chilometri *(frecce in blu)*. Si ritiene che le colonne pliniane siano spesso la prima fase di una catastrofica eruzione destinata a formare una caldera risorgente.

In questo ragionamento sussiste, però, una certa ambiguità: ci si chiede, infatti, in che maniera avvenga il collasso. La pressione esercitata dal plutone potrebbe, ad esempio, indebolire lo strato roccioso che forma il tetto della camera e permettere al magma di sfuggire attraverso una grande frattura anulare sviluppatasi sopra il perimetro del plutone; in questo caso la parte sommitale del vulcano collasserebbe a causa del vuoto venutosi a creare nella camera magmatica per l'estromissione del magma stesso. Un'altra possibilità consiste nell'evenienza che il tetto della camera, a contatto con il magma, ceda in più punti e, sprofondandosi in esso, apra la frattura anulare per la quale il magma può finalmente passare. Secondo T. H. Druitt e R. S. J. Sparks dell'Università di Cambridge, sia nel caso in cui la subsidenza è causa di eruzione, sia nell'altro, in cui ne è invece l'effetto, la frattura anulare, diventata parete della caldera, dovrebbe essere verticale o a imbuto per conciliare la velocità eruttiva e lo sprofondamento del tetto.

Il magma, ricco di vari gas disciolti, e principalmente di vapore acqueo, raggiunge alla sommità del plutone una temperatura generalmente compresa tra i 700 e i 1000 gradi centigradi. Esso risale, dunque, attraverso la frattura anulare di nuova formazione e, nel contempo, subisce una pressione sempre minore fino a che alla profondità di circa 1000 metri i gas si liberano dalla soluzione proprio come succede quando si stappa una bottiglia di champagne. Il magma dacitico o riolitico è, comunque, ben più viscoso dello champagne (come, del resto, anche quello basaltico) e, quindi, sarà anche meno spumeggiante. Inoltre i gas, proprio perché non riescono a sfuggire del tutto, finiscono col frantumare il magma. Il processo è nella realtà piuttosto complicato, ma il risultato è altrettanto semplice: il magma che risale verso l'alto si espande, dà luogo alla formazione di pomici e si frammenta esplosivamente in una moltitudine di particelle solide incandescenti, le cui dimensioni vanno dai micrometri ai metri.

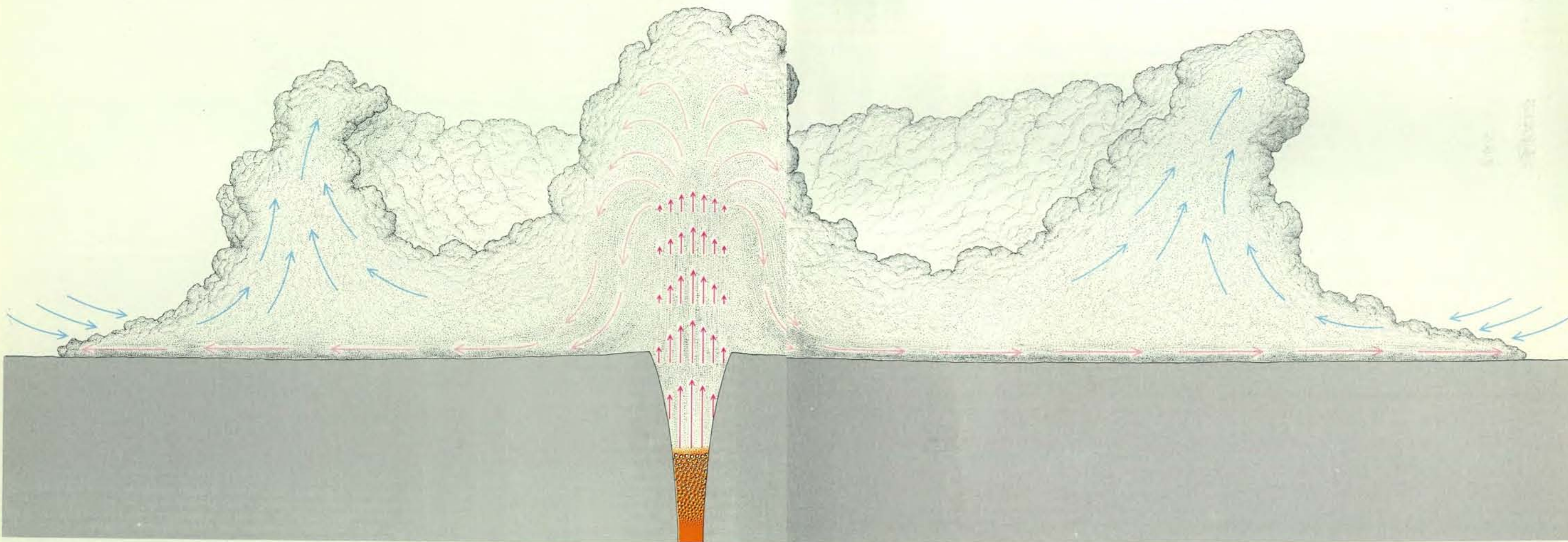
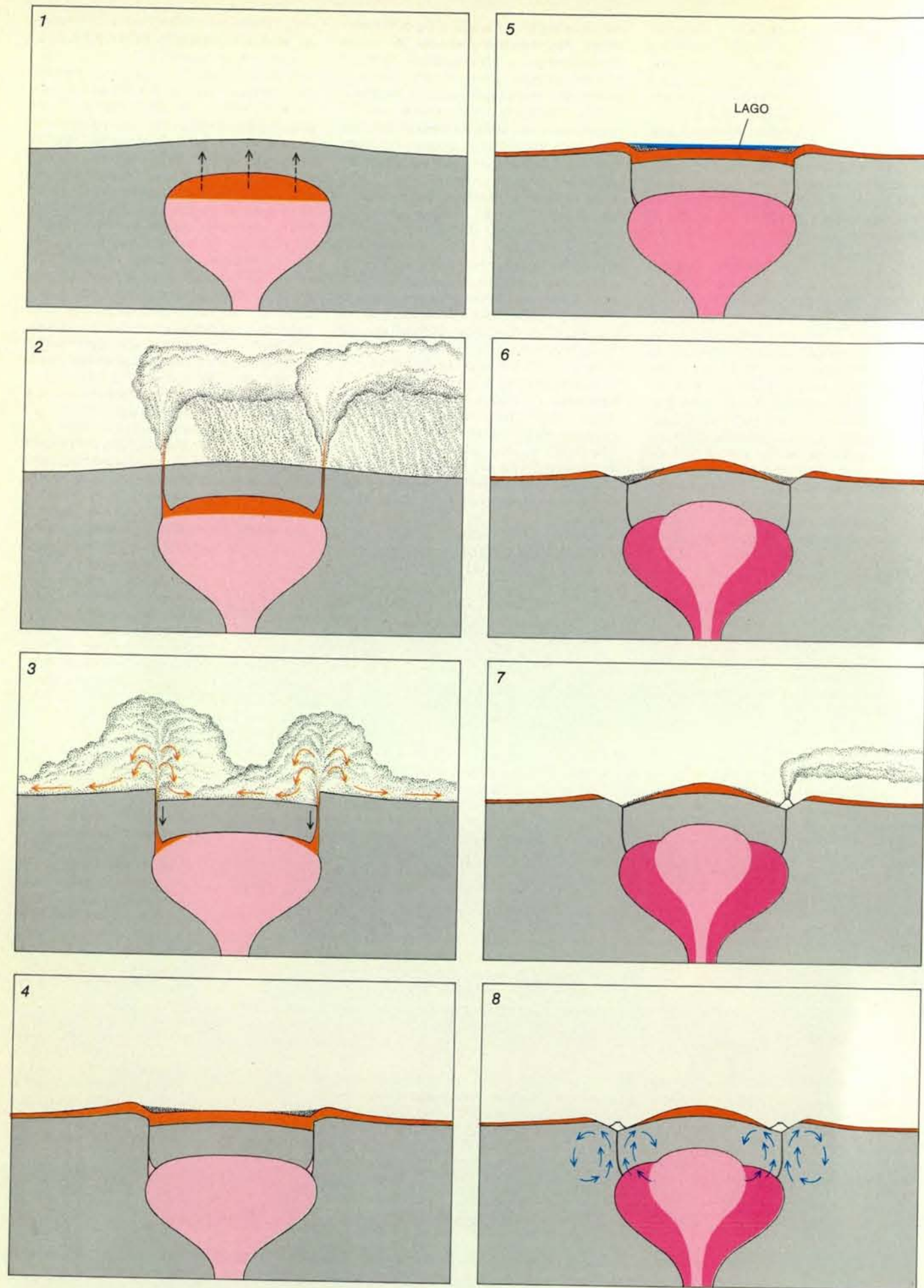
La chiave per comprendere a fondo una grande eruzione che porti alla formazione di una caldera consiste nel chiarire l'attività della miscela incandescente di cenere, polvere, pomici e gas liberi, quando viene scaricata sulla superficie terrestre. La cinetica dell'eruzione è stata studiata nei particolari da un gruppo costituito da G. P. L. Walker dell'Imperial College of Science and Technology, da Sparks dell'Università di Cambridge, da Lionel Wilson dell'Università di Lancaster e da Colin J. N. Wilson dell'Università di Auckland. Essi hanno dimostrato che, quando la velocità dell'eruzione è notevole e la bocca è relativamente piccola (dai 50 ai 100 metri di diametro), si sviluppa una colonna eruttiva che si innalza per decine di chilometri nell'atmosfera. L'esempio più recente è costituito dalla colonna eruttiva del Mount St. Helens, alta circa 20 chilometri (si veda l'articolo *Le eruzioni del Mount St. Helens* di Robert Decker e Barbara Decker in «Le Scienze» n. 153, maggio 1981).

Bisognerebbe sottolineare, però, come le pomici presenti nella colonna non siano spinte fuori dalla bocca come pallini da un fucile da caccia. Infatti, immediatamente al di sopra della bocca, l'energia della colonna è ancora in gran parte cinetica, con velocità pari a diverse centinaia di metri al secondo; in seguito, le pomici rallentano rapidamente e questo non solo per effetto della gravità, ma anche per la resistenza aerodinamica. Ed è a questo punto che un secondo processo comincia a interessare la colonna: tutto l'insieme incandescente di pomici, ceneri e gas risucchia e riscalda l'aria presente intorno alla colonna e si trasforma, così, in una mistura che comincia a risalire per convezione, con nuove accelerazioni verso l'alto. Le colonne eruttive che risalgono per convezione sono ben note, poiché sono la caratteristica principale delle cosiddette eruzioni pliniane, chiamate così dalla celeberrima descrizione della già citata eruzione del Vesuvio, che ci offrì Plinio il Vecchio. La convezione può portare una colonna pliniana ad altezze di anche 50 chilometri.

Imponenti eruzioni pliniane possono segnalare l'inizio di un catastrofico collasso che porta alla formazione di una caldera. Col procedere dell'eruzione, però, le colonne pliniane danno luogo alle nubi ardenti, che costituiscono senz'altro la parte più rilevante dei prodotti eruttivi. Le ragioni sono molteplici: le dimensioni della bocca possono crescere per la violenza iniziale dell'eruzione o per il collasso della caldera, mentre il contenuto in gas del magma può diminuire quando entrano in gioco i livelli inferiori del plutone. Così diventa sempre più improbabile che la colonna eruttiva riesca a sostenersi e, diventando più densa dell'aria che la circonda, comincia ben presto a collassare. Ciò significa che la massa incandescente di gas e pomici finisce con il mantenere solo la prima componente del moto, puramente balistica. Può raggiungere ancora un'altezza di 10 chilometri, ma senza il flusso convettivo non può mantenersi in quota e la colonna ricade spargendo i suoi componenti intorno alla bocca.

Ed è a questo punto che si sviluppano le nubi ardenti. Come già visto, essendo costituite da una miscela di prodotti solidi e di gas, sono dotate di una straordinaria mobilità. Esaminando la distribuzione delle ignimbriti si vede che le nubi ardenti possono risalire lungo rilievi alti anche 1000 metri e percorrere distanze dell'ordine dei 150 chilometri. Per la verità non ne sono mai state osservate a grande scala direttamente, ma il fatto stesso che riescono a risalire lungo rilievi alti anche 1000 metri ci dice che la loro velocità deve essere dell'ordine dei 100 metri al secondo. Le nubi ardenti di minore entità del Mount St. Helens, osservate nelle ultime fasi dell'eruzione, avevano una velocità di 30 metri al secondo.

Sparks e Colin Wilson hanno compiuto studi sia teorici sia sperimentali per spiegare la mobilità delle nubi ardenti. Secondo il loro modello, esse si sposterebbero in condizioni di parziale fluidità. Infatti, i gas che continuano ad abbandonare le caldissime particelle di pomice, combinati con i gas atmosferici che il fronte della nube incorpora, si muovono attraverso le particelle stesse, sfuggendo verso l'alto, e mantengono in sospensione i granuli più piccoli i quali vengono così a costituire un mezzo lubrificato nel quale i clasti più grandi vengono dispersi. La nube viene, quindi, ad avere una minore densità e soprattutto una viscosità molto più bassa di quella che dovrebbe avere una normale aggregazione non fluidificata di pomici e di polvere e, in sostanza, fa un uso molto efficiente della sua energia cinetica che le deriva quando ricade verticalmente da un'altezza di diversi chilometri.

Il principale tipo di deposito di una massiccia nube ardente è una copertura di pomici e particelle minute che, anche a 50 chilometri dalla bocca, può avere uno spessore di parecchi metri. Tutte le minuscole particelle incorporate nella nube possono, inoltre, formare una nube secondaria di cenere, che, per convezione, può risalire per molti chilometri. La cadu-

Una nube ardente si genera quando la velocità dell'eruzione rallenta, quando il magma eruttato non contiene molto gas o quando la bocca eruttiva si allarga. Spesso si forma quando una colonna pliniana collassa, cosicché le particelle risalgono solo balisticamente (*frecce in rosso*) e ricadono, quindi, a terra. È a questo punto che si forma una nube di particelle parzialmente fluida e sostenuta da gas ad altissima temperatura. Questa nube ardente può risalire lungo i fianchi di rilievi alti anche 1000 metri e percorrere distanze di 150 chilometri a velocità di 100 metri al secondo.

ta successiva di particelle da questa nube può dar luogo a un sottile strato di cenere su una regione ben più grande di quella ricoperta dall'ignimbrite formatasi dalla nube ardente. In realtà lo strato, detto di «cenere coignimbritica», può rappresentare anche un terzo dell'intero deposito ignimbritico.

Perché le eruzioni che formano le caldere risorgenti presentano magmi esclusivamente dacitici o riolitici? Si può rispondere con due argomentazioni, tra loro correlate: in primo luogo, i gas presenti in un magma basaltico, a causa della sua minore viscosità, possono facilmente sfuggirne e così l'eruzione finisce con il non essere mai catastroficamente esplosiva e con il non produrre che minime quantità di ceneri; in secondo luogo, proprio perché l'eruzione basaltica è povera di ceneri, il trasferimento di calore dalla colonna eruttiva all'atmosfera è molto piccolo, in quanto esso risulta molto maggiore per le particelle di diametro inferiore al millimetro che non per quelle più grandi. Per questo motivo, è difficile che un'eruzione di magmi basaltici generi correnti convettive nell'atmosfera e diventi una colonna pliniana.

Le «fontane ardenti» osservate nei vulcani attivi delle isole Hawaii costituiscono un ottimo esempio di quanto abbiamo detto. In una fontana ardente, notevoli quantità di magma sono spruzzate verso l'alto, ma per la sua natura basaltica la lava emerge a grandi getti estesi anche più di un metro. Difficilmente viene eruttata cenere fine e i brandelli di lava non trasferiscono direttamente calore all'atmosfera, cosicché non c'è quasi alcuna possibilità che si sviluppi una qualche colonna convettiva. Al contrario, la lava è schizzata tutt'intorno alla bocca. Poiché, quindi, la lava trattiene quasi tutto il suo calore originario, sono abbastanza comuni le colate laviche secondarie.

Uno degli aspetti più sorprendenti delle catastrofiche eruzioni che formano una caldera risorgente è la loro ridotta durata. Michael T. Ledbetter dell'Università della Georgia e Sparks hanno studiato la velocità con cui le particelle di cenere si depositano nell'acqua marina e ne hanno individuato la dipendenza dalla densità e dal raggio di ciascuna particella: pertanto può succedere che particelle minute e altre di più grandi dimensioni, formatesi rispettivamente all'inizio e alla fine di un'eruzione, raggiungano il fondo del mare simultaneamente. Gli stessi Ledbetter e Sparks hanno dedotto dai dati raccolti che i 300 miliardi di metri cubi di ceneri eruttati dalla caldera di Atitlán 84 000 anni orsono (la caldera misura 28 chilometri di diametro ed è priva di centri di risorgenza) si siano depositati in un periodo compreso tra 20 e 27 giorni soltanto e hanno anche escluso che, in questo sia pur breve arco di tempo, l'eruzione sia stata continua. Sembra probabile, invece, che una colonna pliniana sia collassata più volte, deponendo diversi strati ignimbritici l'uno sull'altro. Lo stesso tipo di considerazioni indica che l'eruzione che ha formato la caldera di Toba ha depositato più di 1000 miliardi di metri cubi di cenere in un arco di tempo di appena nove giorni.

All'eruzione seguono diversi eventi. È tipico che un lago riempia la nuova caldera e che sul suo fondo si depositino sedimenti erosi dalle pareti della caldera. Inizia su scala più lenta il sollevamento del fondo della caldera. In generale non si tratta di un semplice sollevamento cilindrico del fondo (e dell'ignimbrite che ora lo ricopre) lungo la frattura anulare. Cartografando l'ignimbrite di Cerro Galán e di altre caldere si vede che essa è inclinata a partire da un punto centrale ben definito, spesso anche di un angolo superiore ai 45 gradi. Il fatto suggerisce che la risorgenza sia un fenomeno locale e che abbia luogo probabilmente sopra un piccolo plutone intrusosi di recente entro il plutone che ha causato l'eruzione. In alcuni casi, tra cui quello di Yellowstone, vi sono in un'unica caldera due centri di risorgenza separati.

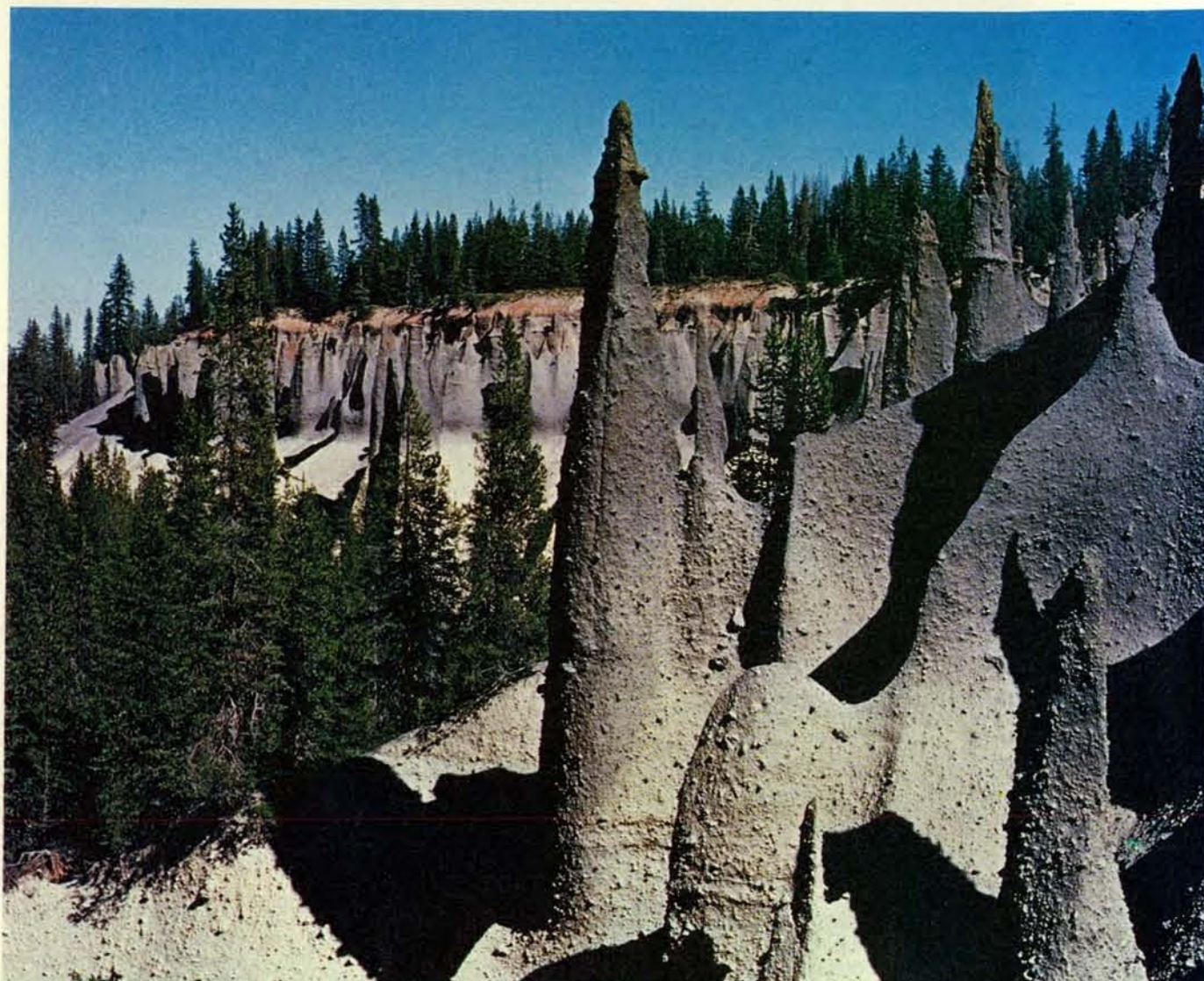
Il riconoscimento che i sedimenti lacustri giovani si erano innalzati per centinaia di metri ha consentito a van Bemmelen di dimostrare che la risorgenza del fondo della caldera di Toba ha formato l'isola di Samosir, la cui superficie è di 640 chilometri quadrati. A Cerro Galán la risorgenza di oltre 1000 metri ha elevato il centro della caldera a oltre 6000 metri di altezza sul livello del mare, dando così origine a una delle più alte montagne dell'Argentina. Poco si sa circa la velocità del sollevamento, ma nel caso della caldera di Long Valley, quella per cui i tempi degli eventi geologici sono meno equivoci, può darsi che il processo sia durato 200 000 anni. Poiché la caldera di Toba ha solo 75 000 anni è probabile che il sollevamento di Samosir sia ancora da completare.

In genere, dopo la risorgenza, interviene, come episodio finale dell'evoluzione di una caldera, un'effusione piuttosto tranquilla di lava dacitica o riolitica che viene emessa da una serie di bocche disposte lungo la frattura anulare. Di solito, la quantità di materiale eruttato è scarsa, ma le effusioni continuano ancora a lungo e a intermittenza. A Long Valley si sono avute tre distinte fasi effusive 500 000, 300 000 e 100 000 anni fa. Si può quindi concludere che gli eventi vulcanici associati alla formazione di una caldera possono continuare, anche se a bassa intensità, per oltre un milione di anni. Le sorgenti termali e i geyser, manifestazioni superficiali della presenza di acque calde che risalgono in superficie dai cosiddetti serbatoi geotermici, possono essere presenti per periodi anche maggiori.

Non si è mai analizzato in dettaglio il rischio che può rappresentare per l'uomo un'eruzione che porti alla formazione di una caldera e questo è accaduto per l'oggettiva difficoltà di fare una simile valutazione. Comunque, molto dipende dal volume totale di emissione e dalle proporzioni della cenere lanciata nell'atmosfera e che poi ricade, dell'ignimbrite e della cenere coignimbritica. Consideriamo, ad esempio, un'eruzione che conduca all'emissione di 1000 miliardi di metri cubi di materiale: un'area di circa 500 chilometri quadrati sprofonderebbe e le ignimbriti riempirebbero interamente la caldera oltre ad altri 30 000 chilometri quadrati di superficie intorno. Lo spessore della copertura sarebbe di oltre 100 metri sull'orlo della caldera e di qualche metro ai bordi della regione interessata. Naturalmente, ogni forma di vita animale e vegetale verrebbe distrutta e le costruzioni sarebbero tutte demolite o sommerse dalla violenza eruttiva.

Nel breve termine la conseguenza più grave dell'eruzione sarebbe, probabilmente, la caduta della cenere coignimbritica, che dovrebbe subire l'influenza dei venti prevalenti oltre che dipendere dalle dimensioni delle particelle. Difatti, le particelle con un diametro di circa 80 micrometri riescono a permanere nell'atmosfera per poche ore al massimo, mentre quelle più piccole vi rimangono più a lungo; ad esempio, le particelle con un diametro inferiore ai 10 micrometri hanno velocità di caduta che non supera quella della turbolenza atmosferica e quindi possono restare nell'atmosfera fino a 100 giorni ed essere trasportate anche a grandi distanze. Le ceneri coignimbritiche sono molto fini e si disperderebbero su un'area di parecchi milioni di chilometri quadrati.

Immaginiamo che l'eruzione del Mount St. Helens, amplificata di un fattore 100, avvenga in questi giorni a Yellowstone. A prescindere dagli effetti

Fasi distinte contraddistinguono l'evoluzione di una caldera risorgente. Nella prima fase, il magma a profondità di 4000-5000 metri forma un plutone, o camera magmatica, che lentamente curva a forma di duomo la superficie *(1)*. Nel plutone vi sono zone di differente composizione, con magma viscoso e ricco di silice (SiO_2) e gas disciolti nella parte alta *(in arancione)*. Quando ha inizio un'eruzione si sviluppano colonne pliniane lungo una grande frattura anulare *(2)*. Di lì a poche ore (o pochi minuti) il tetto del plutone collassa lungo la frattura formando una caldera e dalle colonne pliniane hanno origine le nubi ardenti *(3)*. L'eruzione dura solo pochi giorni. La caldera e anche l'area che la circonda si riempie di ignimbrite *(4)*. Nel frattempo le pareti della caldera cominciano a subire un processo di erosione *(puntini neri)* e si può anche formare un lago *(5)*. Più tardi, in un periodo di centinaia di migliaia di anni, una nuova intrusione di magma nel plutone spinge il fondo della caldera a sollevarsi di nuovo verso l'alto *(6)*, mentre un'attività vulcanica ridotta persiste lungo la frattura anulare *(7)*. Il calore del plutone può dare origine a correnti convettive di acque ricche di minerali *(8)* e dar luogo in superficie a sorgenti termali e geyser, attivi per milioni di anni.

I due strati di cenere visibili al Crater Lake, nell'Oregon, derivano da un'eruzione di 7000 anni fa che portò dapprima all'emissione dei livelli superiori del plutone, più viscosi, e in seguito a quelli inferiori, che lo erano meno. Gli strati di cenere ribaltano questa relazione temporale. Lo strato inferiore, di colore chiaro, fu il primo a essere depositato ed è ricco in silice e povero in ferro e magnesio; lo strato superiore, di colore più scuro, venne depositato immediatamente dopo ed è ricco in ferro e magnesio e povero, al contrario, in silice. Il diametro del Crater Lake misura soltanto nove chilometri ed è quindi troppo piccolo per poter essere considerato alla stregua di una caldera vera e propria.

immediati, tra i quali il blocco repentino di qualunque forma di traffico aereo o terrestre da parte della cenere coignimbritica, le conseguenze più pesanti ricadrebbero, probabilmente, sull'agricoltura, dal momento che milioni di chilometri quadrati di terreni coltivati a cereali verrebbero ricoperti da parecchi centimetri di cenere. È facile prevedere che il raccolto di un anno e forse di più anni andrebbe distrutto. Sul lungo termine, poi, il clima della Terra sarebbe influenzato dall'immissione di polvere nell'atmosfera insieme con un persistente aerosol di particelle creato dall'azione dei gas vulcanici acidi. Ne risulterebbe, allora, un danno per l'agricoltura mondiale che potrebbe durare diversi anni. Si potrebbe prevedere un'eruzione di questo tipo? L'indicazione più verosimile potrebbe essere la fuoriuscita in superficie di magma dacitico o riolitico man mano che si forma una frattura anulare. Contemporaneamente, segnali sismici potrebbero indicare il movimento di magma in una camera situata a una profondità di alcune migliaia di metri. (Dei 2 tipi di onde sismiche, *P* ed *S*, le *S* sono le più interessanti, perché non si trasmettono attraverso i liquidi.) La posizione della camera dovrebbe essere poi confermata da una anomalia locale del campo gravitazionale terrestre, essendo la densità del magma inferiore a quella delle rocce crostali solide. Un ulteriore segnale indicatore potrebbe essere un certo rigonfiamento, o sollevamento, del terreno e in particolare un aumento della velocità di innalzamento che le tecniche convenzionali di sorveglianza sarebbero in grado di rilevare facilmente.

La difficoltà consiste nel sapere interpretare tutti questi fenomeni come segni premonitori di un dato evento e formulare una previsione temporale in assenza di casi storici. Attività sismica e rigonfiamento sono stati recentemente osservati nella caldera di Long Valley e lì si è intesi come indicatori della presenza di nuovo magma alla profondità di parecchie migliaia di metri. Ci si chiede, allora, se a questi eventi seguirà una grande eruzione o se essi precedano semplicemente una tranquilla estrusione in superficie di un duomo lavico. Purtroppo, i segni che immediatamente precedono un'eruzione che possa generare una caldera non sono ben conosciuti e, per di più, l'attività che fa presagire l'evento può durare anche per molte generazioni. Può darsi non sia possibile, quindi, prevedere un'eruzione con esattezza. Ciononostante non si dovrebbero trascurare i segni di una rinnovata attività in qualsiasi caldera.

Anche se può sembrare paradossale, le eruzioni che danno origine a caldere risorgenti sono così rare che, tutto sommato, possono essere considerate persino vantaggiose per l'uomo. Infatti, a una caldera che si trova sopra un plutone è sempre associata una anomalia termica della crosta, che può continuare a manifestarsi in sorgenti termali anche per due o tre milioni di anni dopo l'eruzione. A parte lo splendido spettacolo offerto dai geyser, l'acqua e il vapore caldo costituiscono potenziali sorgenti di energia utile, e così le rocce calde che ricoprono il plutone. Negli Stati Uniti si sta attuando un importante esperimento di tecniche geotermiche con l'Hot Dry Rock Project del Los Alamos National Laboratory. Si pompa acqua in un pozzo piuttosto profondo situato lungo i fianchi della caldera di Valles, per portarla in una zona di rocce calde e fratturate. L'acqua, così riscaldata, ritorna in superficie attraverso un secondo pozzo e può essere utilizzata per generare elettricità.

Alcuni benefici, forse meno apparíscenti, legati alla presenza di caldere, consistono nei depositi di minerali che si sono formati nelle caldere risorgenti stesse, o nelle loro vicinanze. La caldera di Kari Kari costituisce un esempio tipico dal punto di vista geologico e ricco d'associazioni storiche. Lungo i suoi fianchi, infatti, sorgono la città di Potosí e il Cerro Rico, un picco di 4900 metri composto da riolite che venne intrusa nella frattura anulare della caldera circa 13 milioni di anni fa. Man mano che raffreddava la riolite venne trasformata chimicamente dalle soluzioni idrotermali: liquidi caldi e acidi giunti in superficie dopo le reazioni chimiche intercorse tra l'acqua e le rocce profonde molto calde. Nel XVII secolo Potosí era la maggiore città dell'emisfero occidentale, proprio per l'abbondanza dell'argento estratto dai giacimenti depositatisi dalle soluzioni termali. L'importanza di questo centro minerario per la Spagna è stata enorme; a essa si deve il celebre detto spagnolo: «vale quanto Potosí». Ma alla fine, il flusso di argento proveniente da Potosí e da altre miniere del Nuovo Mondo provocò una disastrosa inflazione della moneta spagnola. D'altra parte, poiché lo sfruttamento delle vene d'argento del Cerro Rico è stato pressoché continuo a partire dal 1544, le miniere si sono alla lunga esaurite mentre ai livelli più profondi si comincia a cercare lo stagno.

Naturalmente, le profonde sorgenti di calore che creano le caldere sono in grado di conservare la circolazione idrotermale per lunghi periodi e l'attività idrotermale può dar luogo a reazioni chimiche con conseguente precipitazione di minerali metalliferi. Può darsi passi del tempo prima di arrivare a comprendere tutti i dettagli del processo, ma il fatto che le caldere siano associate all'energia geotermica e a giacimenti minerari di valore può essere vantaggioso e non farci apparire le caldere risorgenti solo come catastrofi.

L'uomo di Pechino

Vissuto per 230 000 anni in caverna, ha lasciato una grande abbondanza di fossili e manufatti, che oggi forniscono informazioni sulla sua evoluzione biologica, tecnologica e forse anche sociale

di Wu Ju-kang e Lin Sheng-lun

Uno fra i siti che possono vantare la più lunga storia di insediamenti da parte dell'uomo o di suoi progenitori è una caverna in prossimità della stazione ferroviaria di Choukou-t'ien, una cittadina che si trova a una cinquantina di chilometri a sud-ovest di Pechino. Mentre le città moderne non hanno in generale più di alcune migliaia di anni di storia, la caverna di Choukout'ien fu occupata quasi di continuo per più di 200 000 anni. I depositi fossiliferi in strati multipli indicano che antichi gruppi umani cominciarono ad abitare la caverna 460 000 anni fa, e gli ultimi di loro abbandonarono il sito solo 230 000 anni fa, quando furono costretti a rinunciare alla loro abitazione in quanto la caverna si era ormai riempita di detriti e di sedimenti. La specie che visse nella caverna è classificata come *Homo erectus pekinensis*, o uomo di Pechino.

La lunga documentazione di questa permanenza umana a Choukout'ien offre un'opportunità di ricostruire lo sviluppo di una singola comunità su un periodo di tempo che copre una frazione importante dell'evoluzione del genere *Homo*. Il periodo è abbastanza lungo da consentire di discernere mutamenti progressivi nella morfologia dei fossili stessi e uno dei mutamenti importanti, riscontrati nei caratteri fisici, è un aumento della capacità cranica. Fatto altrettanto ragguardevole, si possono ricostruire certi eventi dell'evoluzione culturale della specie. Nei depositi della caverna sono stati trovati indizi del fatto che l'uomo di Pechino era in grado di avere un controllo sul fuoco, che utilizzava per cuocere i cibi. I resti fossilizzati di ossa animali indicano che gli abitanti della caverna di Choukout'ien erano abili cacciatori di prede grandi e piccole, mentre semi fossilizzati suggeriscono un'altra componente della dieta. Una grande quantità di manufatti di pietra scheggiata forniscono informazioni sull'abilità di questi uomini nella lavorazione della pietra. Analizzando i materiali recuperati nella caverna è possibile anche avanzare delle congetture sull'organizzazione sociale della comunità: per esempio ci sono indizi a sostegno dell'ipotesi che questi gruppi praticassero la spartizione del cibo e la divisione del lavoro secondo i sessi.

Negli ultimi cinque anni una vasta ricerca sul sito di Choukout'ien è stata intrapresa da più di 120 scienziati cinesi, fra cui noi due. Essa è patrocinata dall'Istituto di paleontologia dei vertebrati e di paleoantropologia dell'Accademia delle scienze cinese. Gli scienziati che vi partecipano rappresentano 17 università e istituzioni di ricerca e i loro campi di specializzazione si estendono dall'uomo preistorico e dai suoi manufatti allo studio del clima in ere antiche, delle caverne e dei terreni in cui esse si formano, degli antichi suoli e del polline, e alla datazione di materiali archeologici. Presenteremo qui le scoperte salienti compiute da questo gruppo e tenteremo di abbozzare un ritratto dell'uomo di Pechino nel suo ambiente nativo.

Nello schema complessivo dell'evoluzione dei primati superiori, l'uomo di Pechino e gli altri esempi di *Homo erectus* rappresentano uno sviluppo relativamente tardivo. Il primo ominoido, progenitore dell'uomo e delle scimmie antropomorfe, si separò dagli altri primati circa 35 milioni di anni fa. Era un quadrupede che si cibava di semi e di frutta. Da 10 a 8 milioni di anni fa il primo ominide, fondatore di una linea genetica il cui unico rappresentante vivente è l'uomo moderno, si separò dagli altri ominoidei come lo scimpanzé e il gorilla. Gli ominidi più primitivi avevano a quanto pare un cranio piccolo, della capacità di forse 350 centimetri cubi; erano in grado di camminare eretti e forse riuscivano a trasportare oggetti mentre camminavano. Solo quattro milioni di anni fa circa comparve un ominide progredito, che viene classificato nel genere *Australopithecus*. I più antichi resti di australopitecini sono stati trovati nella valle del medio corso del fiume Auàsc, in Etiopia. Lo scheletro era evidentemente capace di sostenere una deambulazione pienamente bipede e la scatola cranica aveva una capacità vicina ai 500 centimetri cubi.

Negli ultimi 100 anni molti fossili di crani simili a quelli dell'uomo moderno e di frammenti di scheletri più evoluti di quanto non siano quelli degli australopitecini sono stati trovati in Europa, Asia e Africa. Per qualche tempo ogni scoperta dette quasi invariabilmente origine a una nuova divisione tassonomica, ma in seguito la massima parte dei nuovi esemplari trovati è stata raggruppata nella singola specie *Homo erectus*. Sulla base di studi cronologici sul contesto geologico in cui furono trovati i vari fossili si è accertato che questa specie emerse almeno 1,5 milioni di anni fa. L'*Homo erectus* era in grado di camminare eretto (come suggerisce il suo nome), ma aveva un cranio ancora molto primitivo, con una capacità variabile fra 850 e 1100 centimetri cubi. Egli sapeva fabbricare utensili in pietra più progrediti rispetto ai *pebble tools*, gli utensili su ciottolo, degli ominidi anteriori. Alcuni *Homo erectus* più tardivi, fra cui l'uomo di Pechino, svilupparono una cultura molto complessa, caratterizzata dall'uso di caverne come abitazioni e dalla caccia. Essi sapevano inoltre usare il fuoco per cuocere i cibi, anche se non è chiaro se fossero veramente capaci di accenderlo.

Il corso successivo dello sviluppo dal-l'*Homo erectus* all'uomo moderno non può essere ricostruito nei particolari. Basti dire che 200 000 anni fa circa apparve una forma umana dal volto meno massiccio e dalla scatola cranica più capiente di quella dell'*Homo erectus*. I membri di questa nuova specie, primi rappresentanti dell'*Homo sapiens*, si diffusero in Europa, Asia e Africa un po' meno di 100 000 anni fa: essi erano capaci di costruire ripari all'aperto e di accendere il fuoco, e costruivano utensili litici molto raffinati. Data la stretta somiglianza della loro morfologia con quella dell'uomo moderno e il carattere progressivo della loro cultura, sono considerati fra i progenitori diretti del moderno *Homo sapiens*.

La caverna di Choukout'ien

La caverna di Choukout'ien, dove sono stati trovati i resti dell'uomo di Pechino, si trova in una formazione di calcare nota come Collina delle Ossa di Drago. Prima di diventare famosa come sito abitato dall'uomo preistorico, essa era stata frequentata già da molto tempo dai raccoglitori di ossa di drago. I draghi ebbero un posto importante nella cultura cinese tradizionale e le loro ossa (in realtà fossili di vari mammiferi) erano considerate di grande valore medicinale. Furono di fatto questi raccoglitori a trovare i primi indizi della presenza dell'uomo di Pechino, pur non riconoscendone il significato.

All'inizio del Novecento un certo numero di paleontologi e di antropologi, fra cui il geologo svedese J. G. Anderson e il paleontologo dell'American Museum of Natural History Walter W. Granger, si recarono a Choukout'ien alla ricerca di tracce dell'uomo primitivo. Nel 1921, nel corso di una ricognizione ordinaria alla Collina delle Ossa di Drago, Anderson, Granger e il paleontologo austriaco O. Zdansky, che era assistente di Anderson, furono indirizzati da gente del luogo a una caverna che si diceva fosse «piena di ossa di drago». Non si trovò però niente di interessante, tranne alcuni frammenti di quarzo; poiché questi frammenti erano lontani dal loro contesto geologico, Anderson li considerò manufatti di antichi esseri umani.

Scavi di ricognizione vennero intrapresi nella caverna, ma il progresso delle scoperte fu lento. Nel 1923, Zdansky trovò due denti di tipo antropomorfo. Quando, nel 1927, fu iniziato uno scavo sistematico, il paleontologo svedese Birger Böhlin scoprì un molare inferiore ben conservato. Davidson Black, che era allora professore di anatomia al Collegio medico di Pechino, assegnò questo fossile alla nuova specie *Sinanthropus pekinensis*. Dal 1927 al 1937, quando i lavori furono interrotti dall'invasione giapponese della Cina, gli scavi nella caverna continuarono in modo sistematico. Nell'inverno del 1929, Pei Wen-chung, che diresse gli scavi a Choukou-t'ien dal 1928 sino alla fine del 1935, trovò il primo cranio completo, fornendo in tal modo una solida base allo studio scientifico dell'uomo di Pechino.

Dopo la proclamazione della Repubblica Popolare Cinese, che avvenne nel 1949, gli scavi nella caverna sono stati ripresi a opera dell'Istituto di paleontologia dei vertebrati e di paleoantropologia. Entro il 1966 vennero trovati fossili di più di 40 uomini e donne di varia età, associati a decine di migliaia di manufatti in pietra e a tracce dell'uso del fuoco. Furono recuperati, inoltre, i fossili di due specie di piante (e il polline di molte altre) e di 96 specie di mammiferi. L'abbondanza dei fossili e la diligenza con cui sono stati studiati consentono di ricostruire elementi della storia dell'uomo di Pechino, prolungatasi per oltre 200 000 anni.

La caverna, chiamata formalmente Località I di Choukout'ien, è un'enorme caverna carsica in una struttura calcarea formatasi nell'Ordoviciano, circa 500 milioni di anni fa. Il paesaggio carsico, che prende il nome dal Carso, si forma per

Questa caverna scavata nel calcare fu abitata dall'uomo di Pechino, una varietà di *Homo erectus*, a cominciare da 460 000 anni fa sino a 230 000 anni fa. Essa fu scoperta all'inizio degli anni venti. Scavi condotti negli ultimi 50 anni e più hanno portato in luce tracce dell'uso del fuoco, una collezione di 100 000 manufatti litici, ossa fossilizzate e denti di più di 40 individui e resti fossili di specie animali e vegetali che fecero presumibilmente parte della dieta dell'uomo di Pechino. La fotografia mostra l'ingresso orientale alla caverna, che fu il primo accesso nella fase iniziale di occupazione. Esso fu abbandonato dopo un massiccio crollo avvenuto all'interno della caverna circa 300 000 anni fa.

erosione, a opera di acque sotterranee su grandi masse di calcari solubili. La caverna carsica abitata dall'uomo di Pechino è lunga 140 metri da est a ovest. La parte più larga, che misura 40 metri, si trova all'estremo orientale; la parte più stretta, all'estremo occidentale, misura solo poco più di due metri.

La caverna originaria non era però così, né in principio fu abitabile da alcun ominide. La sua formazione all'interno della collina ebbe inizio nel Pliocene, circa cinque milioni di anni fa, quando l'acqua, filtrando attraverso fessure orizzontali e verticali cominciò a sciogliere il calcare e a trasportarlo via. L'erosione orizzontale creò una lunga caverna con le due parti estreme più strette dell'area centrale; l'erosione verticale seguì fenditure molto lunghe, creando gole a forma di imbuto, dal fondo profondo, aspro e accidentato. La caverna originaria, all'interno della Collina delle Ossa di Drago, non aveva aperture in superficie. Però, all'inizio del Pleistocene, a causa della continua erosione della collina da parte dell'antico fiume Choukou, un contrafforte sul versante orientale della collina stessa si staccò, esponendo così all'erosione da parte dell'acqua la parete orientale della caverna. Di conseguenza, sulla pendice orientale si sviluppò una fenditura diagonale press'a poco in corrispondenza dell'attuale posizione del piccolo ingresso a est della caverna: essa divenne progressivamente più grande e, infine, stabilì una comunicazione fra il fiume Choukou e la caverna. Sabbia e pietrisco furono trasportati dalle acque del fiume all'interno della caverna, livellandone gradualmente il fondo irregolare e creando un pavimento spazioso e piano. Fu a questo punto che l'uomo di Pechino fissò qui il suo domicilio.

Dapprima l'uomo di Pechino stabilì il suo ingresso da est e visse principalmente nella sezione orientale della caverna. Circa 350 000 anni fa, però, ci fu una grande frana nella parte della sezione orientale, oggi chiamata la Sala della Colomba. Una quantità enorme di roccia erosa precipitò dalla volta, bloccando completamente l'ingresso orientale. L'uomo di Pechino dovette abbandonare quella che era stata la parte più spaziosa della caverna, ma poté rientrare nella caverna stessa attraverso una fenditura nella sezione centrale che gli consentì l'accesso alla parte occidentale, dove visse sino a 230 000 anni fa, quando il luogo fu riempito completamente dai detriti dell'occupazione umana e dal calcare eroso.

La caverna rimase piena di questi depositi finché non cominciarono gli scavi. I depositi contenenti i resti dell'uomo di Pechino avevano uno spessore di circa 40 metri, che nel 1929 furono divisi in dieci strati da Pierre Teilhard de Chardin e dal paleontologo cinese Yang Chung-chian. Nel 1951, l'archeologo cinese Chia Lan-po analizzò i depositi al di sotto del decimo strato e definì altri tre strati (dall'11 al 13). I vari ricercatori sono stati concordi nell'identificare questi strati come quelli del periodo di Choukout'ien. È stato generalmente riconosciuto che questo periodo appartiene al Pleistocene medio, cioè risale a circa 500 000 anni fa, ma ancora recentemente la cronologia era così generica e priva di particolari da non fornire alcuna indicazione circa l'evoluzione dello stesso uomo di Pechino.

Di recente, decine di ricercatori, appartenenti a cinque istituti, hanno terminato una serie di ricerche cronologiche, fornendo quindi, per la prima volta, una data relativamente esatta per ciascuno strato. Il decimo strato è stato datato contando le tracce lasciate in minerali cristallini dalla disintegrazione di nuclei di uranio 238, l'isotopo più abbondante dell'uranio. Il tasso di tali eventi di fissione è costante, cosicché il numero delle tracce è una misura del tempo trascorso dopo la formazione del minerale. La conclusione a cui si è giunti è che il decimo strato fu deposto 460 000 anni fa.

Ricercatori dell'Istituto di geologia dell'Accademia delle scienze cinese hanno accertato, con la tecnica di datazione nota come «metodo della famiglia dell'uranio», che il nono e l'ottavo strato furono deposti 420 000 anni fa. In tale metodo ciò che si misura sono le abbondanze relative dell'uranio 234 e del prodotto del suo decadimento radioattivo: il torio 230. Il decadimento procede a una velocità costante, che è nota; il rapporto fra i due elementi indica perciò il tempo trascorso dopo la formazione del minerale uranifero.

Un altro metodo ancora fu applicato alla datazione del settimo strato, la cui età risulta compresa fra 370 000 e 400 000

La caverna si trova nei pressi di Choukout'ien, 50 chilometri circa a sud-ovest di Pechino *(a destra)*. Essa dipende amministrativamente da Fangshan, distretto suburbano di Pechino, a circa 150 chilometri dalla costa *(a sinistra)*. Nella parte nordoccidentale della città c'è la Collina delle Ossa di Drago, così chiamata per i fossili di mammiferi che vi venivano raccolti dalla popolazione locale e ritenuti veramente ossa di drago di grande valore medicinale. La Località I di Choukout'ien, uno dei quattro siti paleoantropologici sulla collina, è la caverna abitata dall'uomo di Pechino.

anni, età stabilita esaminando il magnetismo naturale residuo di minerali presenti nello strato. Quando un minerale magnetico cristallizza, il suo asse magnetico si allinea con la direzione del campo magnetico terrestre. Le periodiche inversioni di questo campo lasciano così una registrazione nelle rocce e da ciò si può risalire all'epoca della deposizione. Il metodo della famiglia dell'uranio è stato applicato anche ai depositi dei tre strati superiori. I risultati indicano un'età assoluta di 230 000 anni.

I resti fossili

I fossili dell'uomo di Pechino, scoperti negli ultimi cinquant'anni, sono stati catalogati nel seguente modo: sei crani completi o quasi e altri 12 frammenti di crani, 15 pezzi di mandibole, 157 denti, tre frammenti di omero, una clavicola, sette frammenti di femore, un frammento di tibia e un osso lunato (un osso del polso a forma di mezzaluna). Essi furono trovati sparpagliati all'interno della caverna in 15 punti. Rappresentano materiali appartenenti a più di 40 individui di sesso maschile e femminile, morti in varie età. Da tutto questo materiale è possibile ricostruire la forma fisica e l'aspetto dell'uomo di Pechino.

Come altri esempi di *Homo erectus* emersi nel Pleistocene medio, l'uomo di Pechino aveva uno scheletro molto simile a quello dell'uomo moderno. La differenza principale consiste nel fatto che le ossa degli arti avevano una parete più spessa e una cavità del midollo più piccola. Il cranio dell'uomo di Pechino differiva invece sostanzialmente da quello dell'uomo moderno. Esso aveva pareti molto più spesse e volta più piatta, arcate sopraorbitarie sporgenti e occipite molto appuntito. La capacità cranica era maggiore di quella dell'*Homo habilis,* una specie che fu prospera 1,8 milioni di anni fa, e più grande anche di quella dell'uomo di Giava, una forma un poco più antica di *Homo erectus,* ma era, invece, molto minore di quella dell'uomo moderno. I denti erano più grandi e più robusti di quelli dell'*Homo sapiens;* avevano inoltre tracce di un collare di smalto, chiamato cingolo, attorno alla corona. Il cingolo è un carattere primitivo dell'uomo preistorico.

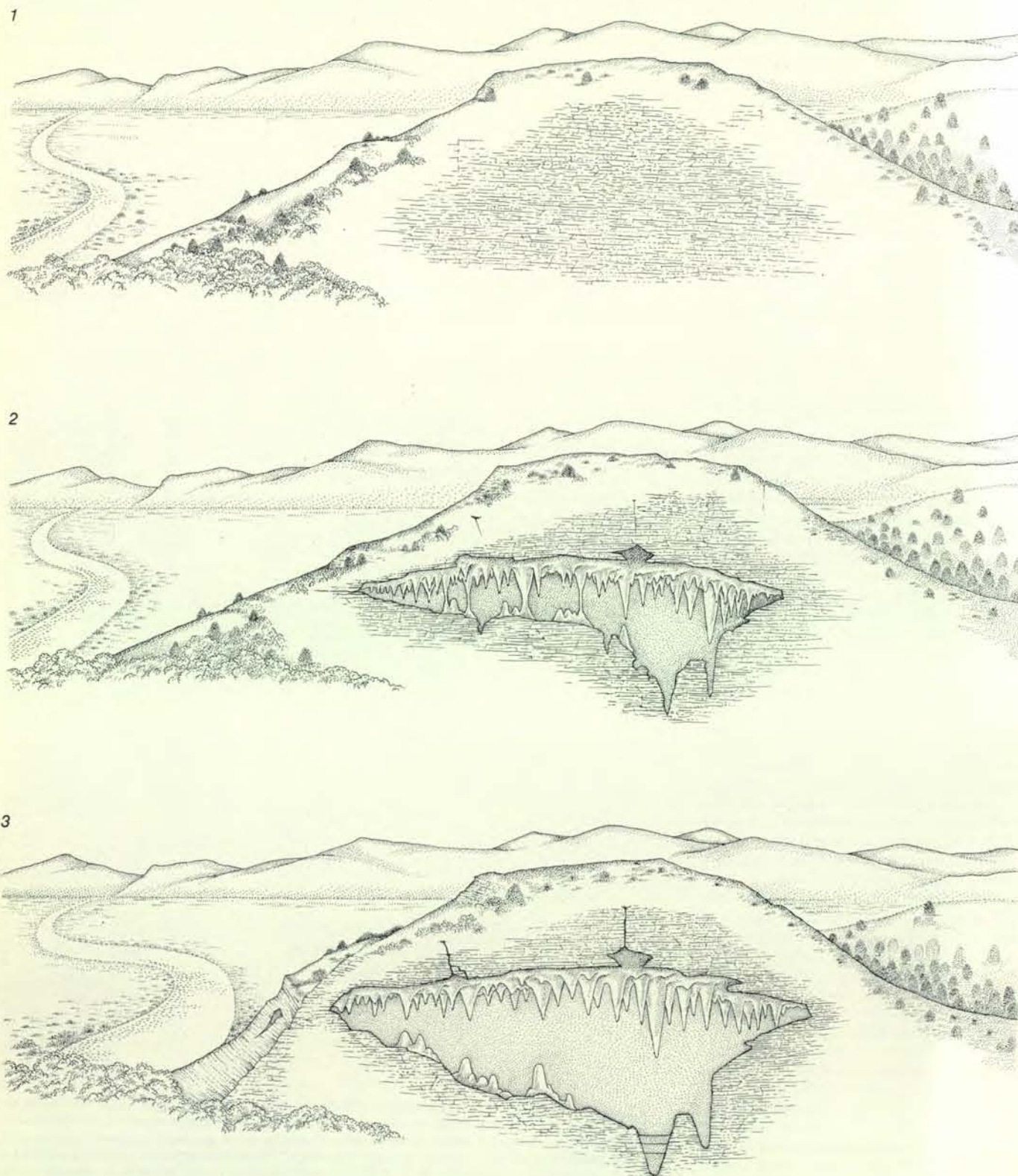
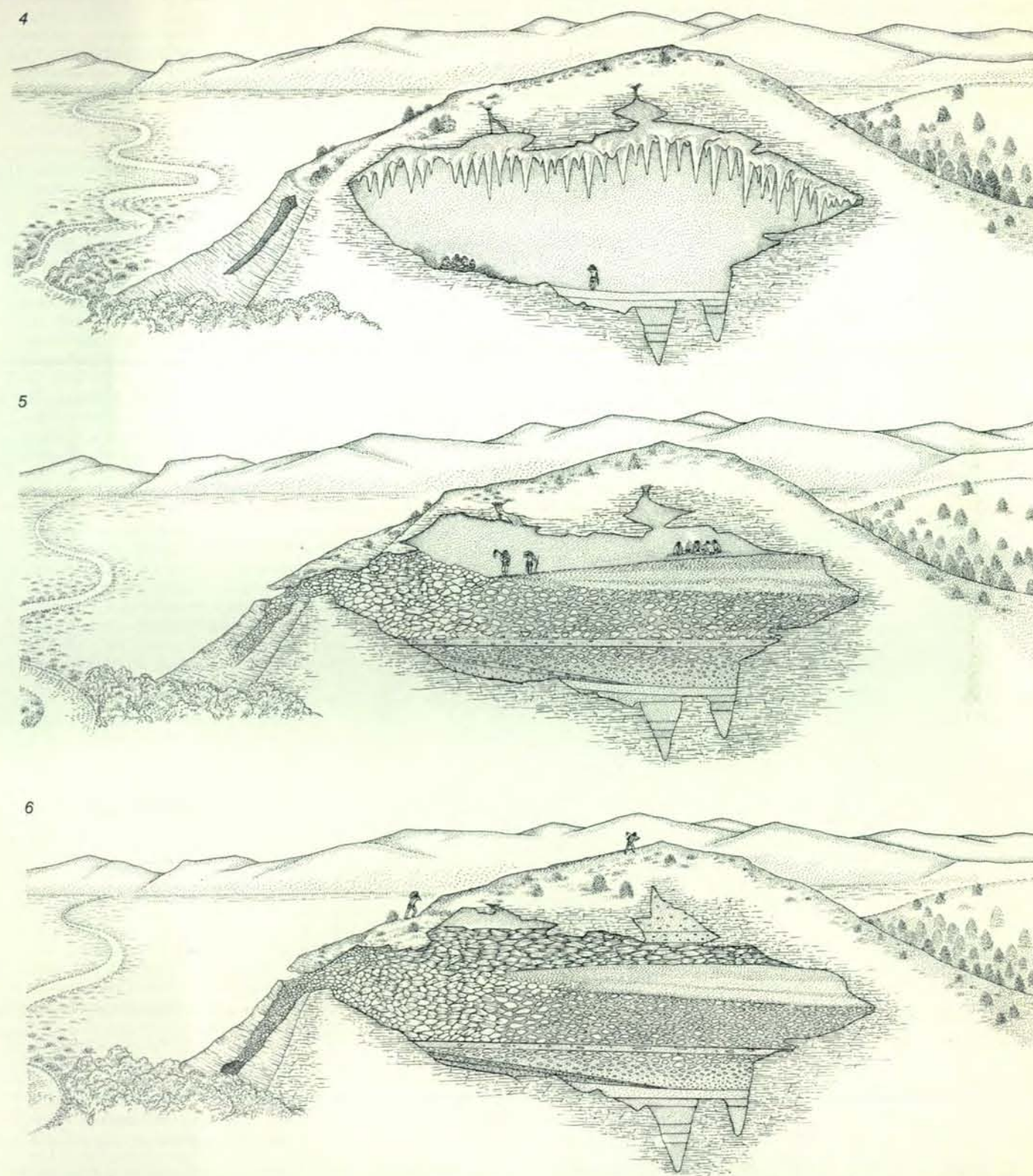
Tutti i fossili dell'uomo di Pechino trovati finora sono stati analizzati da uno di noi (Wu Ju-kang) e da Tung Sing-jen dell'Istituto di paleontologia dei vertebrati e di paleoantropologia. Abbiamo dedicato particolare attenzione alla questione della capacità cranica dell'uomo di Pechino. Va sottolineato che il volume cerebrale non va assunto incondizionatamente come un indicatore dell'intelligenza umana; un ominide con il cervello più voluminoso non è necessariamente più intelligente. Nondimeno l'uomo moderno, con una capacità cranica media di 1450 centimetri cubi, deve essersi evoluto da una forma umana anteriore con cranio di dimensioni inferiori. C'è perciò una tendenza all'aumento nella capacità cranica umana e questa capacità, riferita a una data specie o sottospe-

L'evoluzione umana può essere ricostruita attraverso mutamenti nella morfologia del cranio *(a sinistra)* e tramite l'aumento della capacità cranica *(a destra)*. Il primitivo ominide *Australopithecus,* che fece la sua comparsa circa quattro milioni di anni fa, aveva un cranio dalla volta piatta, con forte prognatismo e piccola capacità cranica (fra 450 e 750 centimetri cubi). Circa un milione e mezzo di anni dopo sopraggiunse l'*Homo erectus*: due esempi di questo ominide, l'uomo di Giava e l'uomo di Pechino, erano relativamente progrediti come struttura facciale e capacità cranica (compresa fra 850 e più di 1000 centimetri cubi); differivano per età, forma del cranio e volume cerebrale. Solo 100 000 anni fa circa comparve l'uomo di Neandertal, primo membro della specie *Homo sapiens*; il suo cranio era molto più grande e meno accentuato il prognatismo. L'uomo moderno, cioè la sottospecie *Homo sapiens sapiens,* emerse all'incirca 40 000 anni fa.

cie, può fornire qualche informazione sul posto che la specie o sottospecie occupa nell'evoluzione umana.

È convinzione generale che, da 3 a 1,5 milioni di anni fa, la capacità cranica sia dell'uomo sia dei suoi progenitori si sia mantenuta stabilmente nell'intervallo fra 500 e 800 centimetri cubi. È questa la capacità misurata in fossili di *Australopithecus* e di *Homo habilis*. L'uomo di Giava, cioè la forma di *Homo erectus* apparsa circa un milione di anni fa, aveva una capacità cranica di 900 centimetri cubi. Quando comparse l'uomo di Pechino, la capacità media aveva raggiunto i 1054 centimetri cubi.

Si dovrebbe sottolineare che questo valore è una media fondata su sei crani relativamente completi dell'uomo di Pechino, trovati a Choukout'ien. Uno dei sei è il cranio di un bambino morto all'età di otto o nove anni. Questo cranio è ovviamente più piccolo di quello di un adulto; non tenendone conto, la capacità cranica media dell'uomo di Pechino salirebbe a 1088 centimetri cubi.

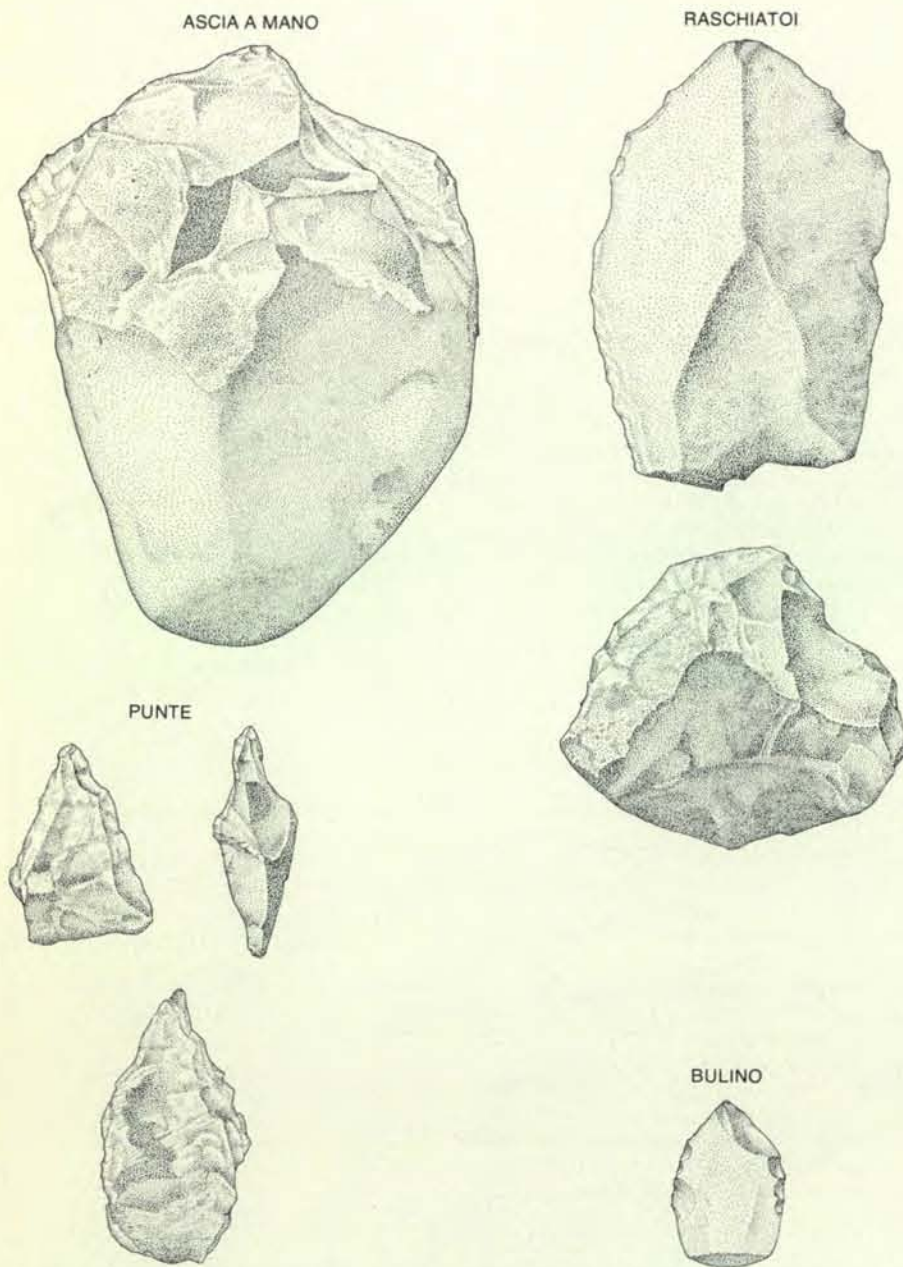
Grazie al lungo arco di tempo coperto dai fossili dell'uomo di Pechino, c'è persino la possibilità di ricostruire dei mutamenti morfologici, in particolare l'aumento della capacità cranica, nel corso della storia dell'insediamento umano in quel sito. La capacità media dei quattro crani trovati nell'ottavo e nono strato (che risalgono a più di 400 000 anni fa) è di 1075 centimetri cubi. L'unico cranio rinvenuto nel terzo strato (230 000 anni fa) ha una capacità di 1140 centimetri cubi. Pare che l'uomo di Pechino, dopo essere vissuto 200 000 anni nella caverna, sia diventato anatomicamente più simile all'uomo moderno.

L'evoluzione della caverna nella Località I di Choukout'ien è schematizzata qui in sei fasi. La Collina delle Ossa di Drago è una struttura calcarea formatasi circa 450 milioni di anni fa *(1)*. Quando circa cinque milioni di anni fa parte del calcare fu disciolta dalle acque sotterranee, la caverna cominciò a formarsi *(2)*. Quando poi, circa tre milioni di anni fa, il contrafforte orientale della collina si staccò in seguito all'erosione da parte dell'antico fiume Choukou, sul versante orientale apparve un piccolo ingresso, che via via si ampliò *(3)*. Sabbia e pietrisco furono trasportati dal fiume all'interno e colmarono i vuoti creati dall'erosione del calcare, livellandone il fondo. La spaziosa caverna fu adottata dall'uomo di Pechino come suo domicilio circa 460 000 anni fa *(4)*. L'ingresso a est e la parte orientale furono le aree più utilizzate come abitazione fino a quando il crollo della volta, 300 000 anni fa, costrinse gli occupanti a passare nella parte occidentale, alla quale si accedeva da una fenditura nella sezione centrale *(5)*. Al termine dell'occupazione, 230 000 anni fa, la caverna era piena di rocce cadute e di detriti *(6)*.

Nella cerchia degli antropologi si riconosce che l'evoluzione morfologica dell'uomo primitivo fu molto più lenta e meno evidente della trasformazione del suo comportamento e stile di vita. Le tecniche usate nella produzione di utensili e gli utensili stessi sono perciò una misura importante dell'evoluzione umana. Nel caso dell'uomo di Pechino, uno studio sistematico della cultura litica fu compiuto nel 1980 da Pei Wen-chung e da Chang Shen-shui dell'Istituto di paleontologia dei vertebrati e di paleoantropologia, i quali sono giunti alla conclusione che l'evoluzione dell'uomo di Pechino si rispecchia in modo più chiaro nei suoi utensili e nel suo modo di produrli che non nei suoi resti fossili.

Gli utensili

L'industria litica dell'uomo di Pechino fu un'industria avanzata, tanto nella scelta dei materiali utilizzati quanto nelle tecniche di lavorazione applicate. Gli utensili in pietra, prodotti dall'uomo di Pechino, sono fatti primariamente di quarzo di vena, cristallo di rocca, selce e arenaria; il che fa pensare che egli non si fondasse esclusivamente su ciottoli arrotondati dall'acqua come materiale per la produzione di utensili. Spesso doveva percorrere distanze piuttosto grandi per procurarsi quarzo di vena e cristallo di rocca, recandosi là dove questi materiali erano rimasti esposti dall'erosione di formazioni granitiche.

La maggior parte degli utensili sono schegge di varie dimensioni conformate secondo una o l'altra di tre tecniche distinte. Nella tecnica detta di percussione su incudine, una grande pietra piatta (l'incudine) era posata al suolo e colpita con forza con un pezzo di arenaria. Le schegge che si staccavano dall'arenaria venivano raccolte, e quelle che avevano una forma e un bordo idonei venivano scelte o per essere usate direttamente o per essere sottoposte a successivi ritocchi. In un'altra tecnica, detta della percussione diretta, il nucleo di selce veniva tenuto in una mano e se ne staccavano schegge colpendolo con un percussore, una pietra che fungeva da martello, tenuta nell'altra mano. La terza tecnica di produzione di schegge si chiama della percussione bipolare. Anche in questo caso un'incudine grande e piatta veniva posta sul suolo: un pezzo di quarzo di vena era tenuto verticale su di essa con una mano e fratturato con colpi verticali, impartitigli con un percussore tenuto nell'altra mano. In questo modo si potevano staccare schegge a entrambe le estremità; i manufatti così ottenuti sono conosciuti con il nome di schegge bipolari.

Alcune schegge recuperate a Choukout'ien venivano usate così com'erano, ma è evidente che alcune furono successivamente trasformate in utensili più specializzati come asce a mano, raschiatoi, punte, bulini e lesine. Il ritocco era eseguito di solito mediante percussione diretta e, nella maggior parte degli utensili, avveniva sulla superficie posteriore anziché su quella che era stata lavorata nell'operazione iniziale di distacco delle schegge. A volte venivano ritoccate entrambe le facce.

Pei e Chang hanno suddiviso l'evoluzione della cultura litica dell'uomo di Pechino in tre periodi basandosi su mutamenti nelle dimensioni degli utensili, nel materiale usato per produrli e nella tecnica di produzione. Gli utensili litici, trovati negli strati compresi fra l'ottavo e l'undicesimo inclusi, rappresentano la cultura più antica, che prevalse fra 460 000 e 420 000 anni or sono e che è caratterizzata da oggetti di ragguardevoli dimensioni, con un peso superiore a 50 grammi e una lunghezza di più di 60 millimetri, oltre che dall'applicazione indiscriminata dei tre processi di produzione. In questa fase, gli utensili fatti di materiali più teneri, come l'arenaria, costituiscono dal 15 al 20 per cento dell'intera produzione. Nella fase intermedia, fra 370 000 e 350 000 anni fa, il processo di percussione su incudine fu praticamente abbandonato e il processo di percussione bipolare divenne il metodo principale per scheggiare le pietre. Di conseguenza salì al 68 per cento la percentuale degli utensili di peso inferiore ai 20 grammi e di lunghezza al di sotto dei 40 millimetri. Di contro gli utensili di grandi dimensioni diminuirono al 12 per cento.

Gli utensili litici prodotti dall'uomo di Pechino sono di quattro tipi principali: l'ascia a mano *(in alto a sinistra)*, il raschiatoio *(in alto a destra)*, la punta *(in basso a sinistra)* e il bulino *(in basso a destra)*. Essi venivano prodotti con quarzo di vena, cristallo di rocca, selce e arenaria. Alcune schegge erano usate come tali, altre venivano ritoccate per dar loro la forma di utensili più specializzati, come punte e bulini. L'industria litica era dominata da piccoli utensili lunghi meno di 40 millimetri e del peso inferiore a 20 grammi, anche se gli utensili primitivi erano in genere più grandi.

L'uso del fuoco da parte dell'uomo di Pechino, attestato da strati di ceneri, può essere ricondotto all'inizio della sua residenza nella caverna. Nei depositi sono stati identificati quattro strati di ceneri: lo strato più consistente ha lo spessore di sei metri, quello più sottile di poco più di un metro.

L'ultima fase durò da 300 000 a 230 000 anni fa. Essa fu chiaramente la più evoluta: gli utensili erano diventati più piccoli e i materiali con cui venivano prodotti erano di qualità migliore. Fra gli utensili di pietra scavati negli strati dal primo al quinto compresi, quelli piccoli salirono addirittura al 78 per cento del totale, mentre quelli di grandi dimensioni scesero ulteriormente al 5 per cento. Gli utensili di questo terzo periodo erano ancora principalmente costituiti da quarzo, ma era diminuito il numero di quelli ottenuti da varietà grossolane come il quarzo di vena; la frazione di utensili in selce aumentò sino a toccare, negli strati superiori, il 30 per cento. Nel frattempo gli utensili di arenaria erano diminuiti a meno dell'1 per cento del totale.

Un'altra indicazione del livello avanzato della cultura dell'uomo di Pechino va vista nell'uso del fuoco: evidentemente egli padroneggiava già quest'arte da quando cominciò ad abitare a Choukout'ien. Nella caverna sono stati trovati quattro grandi e spessi strati di depositi di cenere, risalenti a periodi compresi fra 460 000 e 230 000 anni fa. In certi punti lo strato maggiore, formatosi fra 310 000 e 290 000 anni fa, ha uno spessore di sei metri. Alcuni depositi di cenere si trovano in mucchi disseminati qua e là, il che fa pensare che l'uomo di Pechino avesse la capacità di controllare il fuoco e di tenerlo acceso per molto tempo.

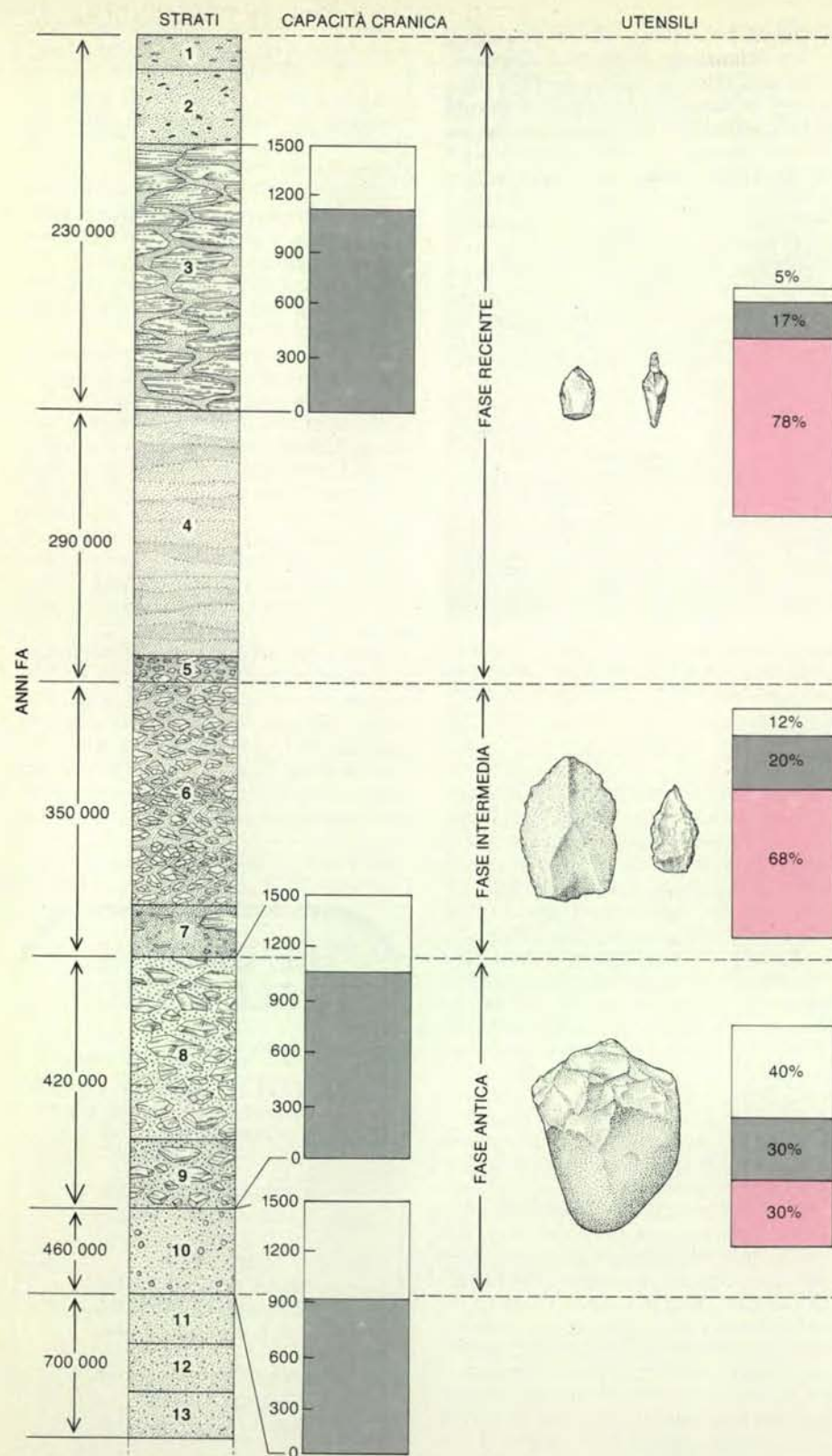
In che modo l'uomo di Pechino accendeva un fuoco e poi lo controllava? Non siamo in grado di rispondere a queste domande in modo decisivo, ma possiamo formulare un'ipotesi. Pare che egli fosse troppo primitivo per avere la capacità di accendere un fuoco e doveva forse dipendere da fuochi naturali prodotti all'esterno della caverna probabilmente da fulmini. Può darsi che catturasse questi fuochi accendendo un fascio di sterpi o piccoli pezzi di legno, che poi portava nella caverna, ma data la loro rarità doveva essere per lui di importanza vitale evitare che la fiamma si spegnesse. Un modo per mantenere acceso un fuoco doveva essere quello di alimentarlo di continuo con legna; un altro quello di coprire il carbone di legna acceso con cenere o terriccio per rallentarne la combustione; in questo caso si poteva riaccendere il fuoco soffiando sui pezzi di carbone. I residui di carbone di legna trovati nella caverna potrebbero essere un'indicazione di questo secondo processo.

Qual era la natura dell'ambiente all'esterno della caverna all'epoca in cui l'uomo di Pechino occupò questo sito? Sulla base di uno studio dei pollini nella Località I di Choukout'ien, Kung Chao-ch'en, dell'Istituto di botanica dell'Accademia delle scienze cinese ha avanzato l'ipotesi, assieme ai suoi colleghi, che l'uomo di Pechino sia vissuto nel clima tipico di un periodo interglaciale: in effetti, il clima di quel periodo non doveva essere molto diverso da quello della Cina settentrionale di oggi. La vegetazione era formata da foreste decidue temperate e da steppe nelle pianure e nelle valli, e da foreste di conifere sulle montagne.

Un clima temperato comporta ogni anno i rigori di un inverno freddo, ma fornisce anche molti tipi di prodotti vegetali commestibili. L'adozione di una caverna come rifugio e l'uso del fuoco potrebbero essere stati ispirati direttamente dal bisogno di sopravvivere in un simile clima.

Numerose piante fornivano non solo legna da ardere, ma anche frutti e semi. Fra i depositi contenuti nella caverna sono state trovate quantità di semi carbonizzati di bagolaro cinese, evidentemente raccolti e arrostiti dall'uomo di Pechino che doveva usarli come cibo. L'analisi dei pollini contenuti nei depositi documenta la presenza di altre piante, come il noce, il nocciolo, il pino, l'olmo e la rosa rampicante. Anche i frutti e i semi di queste specie devono essere considerati possibili voci appartenenti alla dieta dell'uomo di Pechino.

La caccia era una preziosa forma di adattamento all'ambiente, dal momento che la carne poteva fornire più calorie e proteine di una dieta vegetariana. L'uomo di Pechino era evidentemente in grado di competere con successo - come cacciatore - con grandi animali carnivori. L'abbondanza di ossa fossili di mammiferi di varie dimensioni trovate nella caverna indica che l'uomo di Pechino non cacciava solo piccole prede, ma era in grado anche di uccidere animali di grossa taglia.

Fra i carnivori esiste generalmente una relazione naturale fra dimensioni del predatore e dimensioni della preda: una volpe non può uccidere una zebra, la quale è invece il cibo preferito dai leoni. Grazie all'uso di armi, l'uomo di Pechino era in grado di sottrarsi ai limiti intrinseci alle dimensioni corporee. In particolare

L'evoluzione dell'uomo di Pechino durante i 230 000 anni di sua residenza nella caverna di Choukout'ien è suggerita dall'aumento della capacità cranica *(a sinistra)* e dallo sviluppo di una cultura litica più raffinata *(a destra)*. I depositi nella caverna contenenti fossili di questo ominide si suddividono in 13 strati. Sono stati recuperati crani dallo strato 10, dalle zone di confine fra gli strati 9 e 8 e dallo strato 3. Le capacità craniche misurate sono di 915 centimetri cubi per il cranio più antico, una media di 1075 centimetri cubi per quattro crani posteriori e 1140 centimetri cubi per il cranio più recente. Pare che le dimensioni craniche siano aumentate più di 100 centimetri cubi durante l'occupazione della caverna. Lo sviluppo culturale dell'uomo di Pechino si suddivide in tre fasi. Nella fase più antica i manufatti erano principalmente asce a mano e raschiatoi; il 40 per cento era costituito da utensili di grandi dimensioni *(in bianco)*, il 30 per cento di dimensioni intermedie *(in grigio)* e il 30 per cento di piccole dimensioni *(in colore)*. Nella fase più recente comparvero utensili più complessi: la percentuale degli utensili grandi e medi scese rispettivamente al 5 e al 17 per cento, mentre la percentuale di quelli piccoli aumentò al 78 per cento.

egli divenne un efficiente cacciatore di cervi, alcuni dei quali erano più grandi e più veloci di lui. La grande quantità di fossili appartenenti ad almeno 3000 individui di due specie di grossi cervi, *Megaceros pachyosteus*, cervo gigante dalla poderosa mandibola, e *Pseudaxis grayi*, affine al sika, potrebbe essere considerata un'indicazione del fatto che i cervi dovevano essere la preda più comune dell'uomo di Pechino.

Gli adattamenti sociali

Come abbiamo visto, l'uomo di Pechino abitava in caverna, usava il fuoco, cacciava i cervi, raccoglieva semi e sapeva costruire utensili specializzati. Fossili e manufatti attestano il suo adattamento biologico e tecnologico. Più difficile è ricostruire il suo adattamento sociale, essendo disponibili in proposito ben poche indicazioni. Tre ipotesi possono essere nondimeno formulate sulla base dei nostri studi.

In generale la raccolta di prodotti vegetali è un lavoro semplice che può essere intrapreso da singoli individui. La caccia, al contrario, e particolarmente la caccia di grossi animali, è così complessa, difficile e rischiosa da richiedere la cooperazione di numerosi individui. Se ne può dedurre, pertanto, che l'uomo di Pechino, quando cominciò a cacciare i cervi, doveva vivere probabilmente in gruppo piuttosto che in solitudine. Migliaia di fossili di animali, trovati nella caverna, suggeriscono, inoltre, che questi cacciatori primitivi preferivano forse portare le loro prede nella caverna e ivi procedere alla spartizione del cibo con altri membri della comunità, piuttosto che consumarle sul luogo dell'uccisione.

In secondo luogo, la caccia di grossi e veloci animali come il cervo doveva essere un'attività difficilmente praticabile dalle donne, in conseguenza dei loro limiti fisiologici (gravidanza e allevamento dei figli). Si può quindi congetturare che l'attività di caccia dell'uomo di Pechino abbia forse provocato la suddivisione sessuale del lavoro all'interno del gruppo, o abbia almeno contribuito a essa. Può darsi perciò che a quell'epoca si fosse già affermato il modello della suddivisione del lavoro, comune alle società attuali di cacciatori-raccoglitori, secondo il quale la caccia è un'attività maschile e la raccolta di cibi vegetali è un'attività femminile.

In terzo luogo, l'esistenza di una cultura litica in costante progresso, estesa su un periodo di 200 000 anni o più, fa pensare che, proprio nella caverna dell'uomo di Pechino, potrebbe essere stata messa in pratica la forma più antica di istruzione. È fuori discussione che ogni generazione poté ricapitolare l'intera storia dello sviluppo degli utensili litici, dalla percussione di un ciottolo allo scopo di ricavarne un rozzo strumento, ai ritocchi apportati a una scheggia per trasformarla in un raschiatoio specializzato. Le tecniche per la produzione di utensili, come la scienza e la tecnologia moderna, dovevano essere trasmesse dagli anziani ai giovani.

La materia oscura nelle galassie a spirale

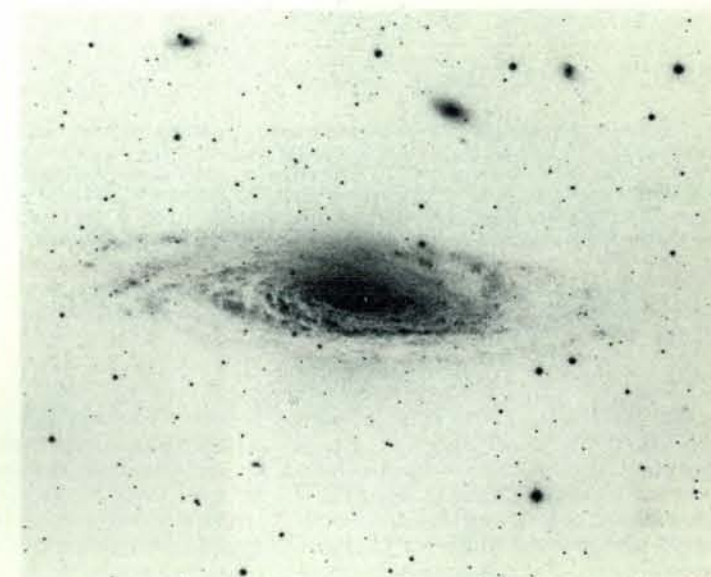
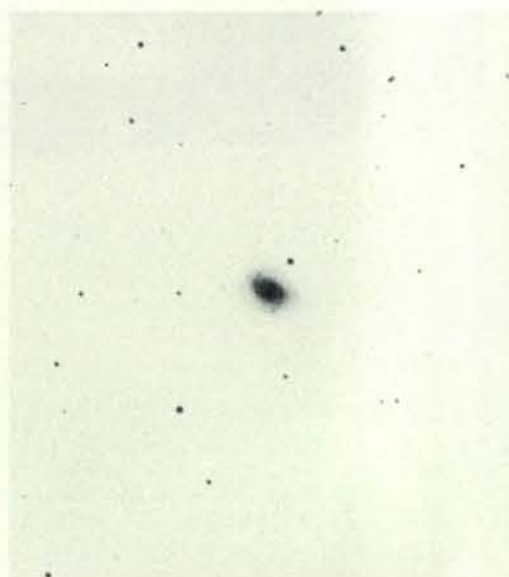
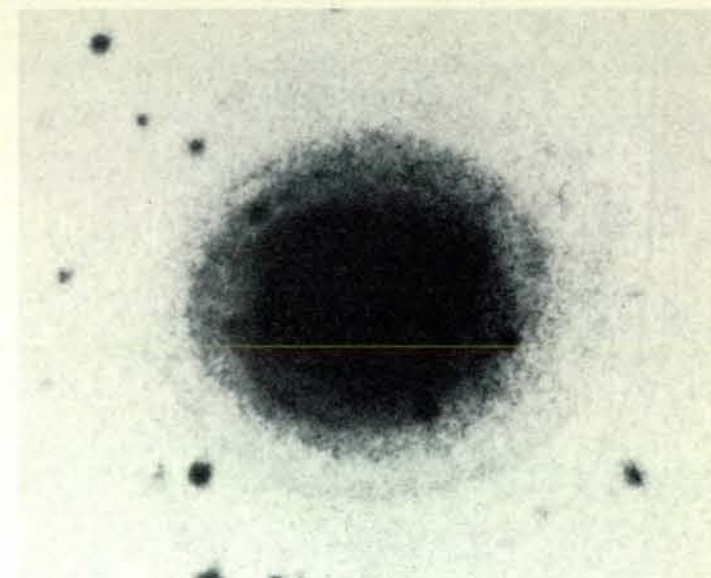
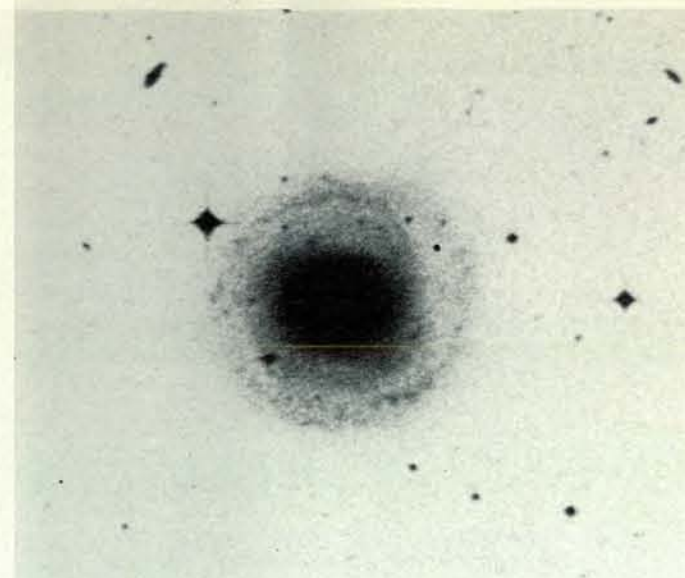
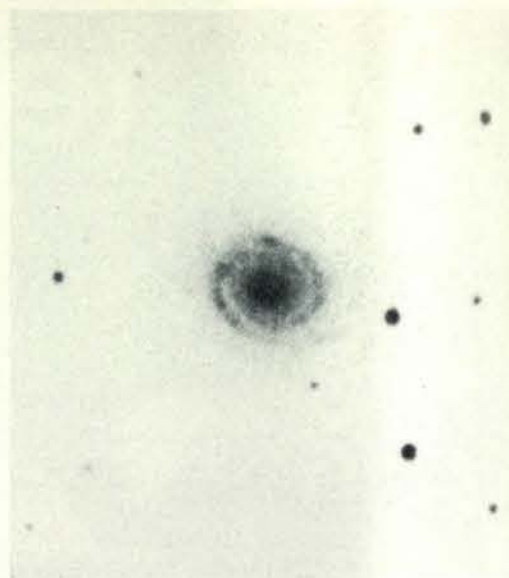
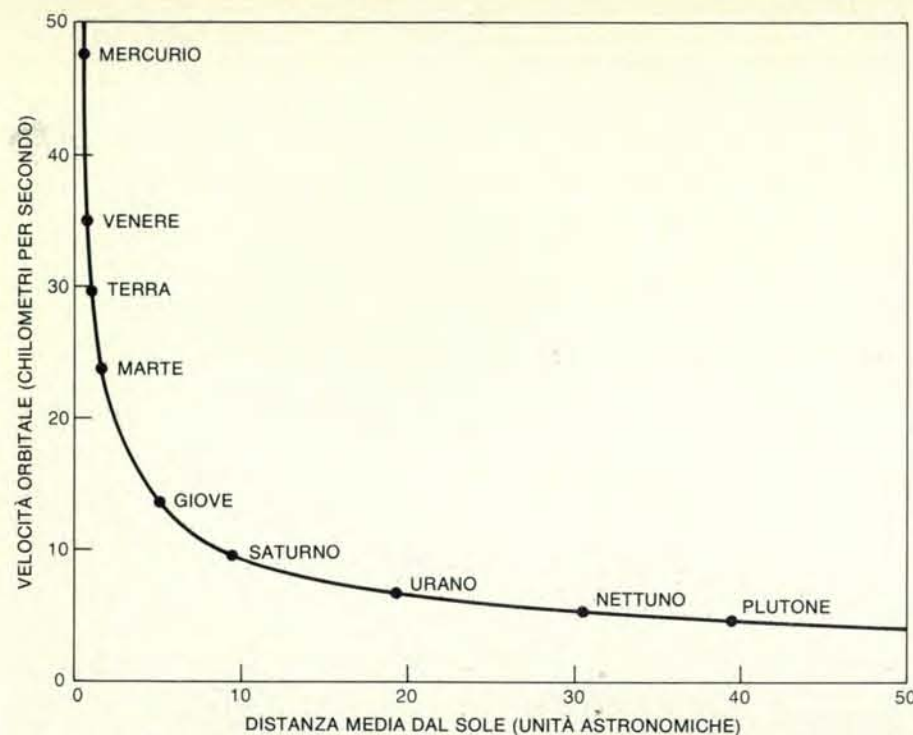
Gran parte della materia contenuta nelle galassie a spirale non è luminosa; inoltre non si trova concentrata vicino al centro della galassia

di Vera C. Rubin

È dalla scoperta delle prime prove osservative dell'espansione dell'universo, avvenuta negli anni venti, che ha senso domandarsi: l'universo continuerà a espandersi indefinitivamente, oppure contiene abbastanza massa perché l'interazione tra gli elementi che lo costituiscono possa rallentare il processo di espansione e alla fine arrestarlo? Oggi quasi tutti i cosmologi sono concordi nel ritenere che l'universo abbia avuto inizio tra i 10 e i 20 miliardi di anni fa in una grande esplosione, il *big bang*, a partire da uno stato infinitamente piccolo e denso, e che da allora abbia sempre continuato a espandersi. Si calcola che la densità critica di materia, necessaria per frenare l'espansione e «chiudere» l'universo, sia dell'ordine dei 50×10^{-30} grammi per centimetro cubo, pari a circa tre atomi di idrogeno per metro cubo. La quantità di materia luminosa presente sotto forma di galassie, però, ammonta solo a $7{,}5 \times 10^{-32}$ grammi per centimetro cubo circa. Per porre, quindi, fine all'espansione dell'universo è necessario che la densità della materia invisibile superi quella della materia visibile di circa 70 volte.

È con questo coefficiente di proporzionalità in mente che gli astronomi continuano ormai da mezzo secolo a sforzarsi di determinare la massa delle galassie che popolano l'universo fino agli estremi limiti raggiungibili dall'osservazione. Dalla luminosità delle galassie tipiche è possibile valutarne la massa, che risulta compresa tra qualche miliardo e qualche bilione di masse solari. In realtà, però, è chiaro che la popolazione stellare di una galassia qualsiasi è molto assortita: vi sono stelle 10 000 volte più luminose del Sole per unità di massa e stelle con una luminosità specifica pari solo a una minima frazione di quella solare. In questa situazione è naturale domandarsi: nelle galassie la distribuzione di luminosità è davvero un indice affidabile della distribuzione di massa? E ancora, per estrapolazione: si tratta di un indice affidabile della distribuzione della massa in tutto l'universo?

Con i miei colleghi del Department of Terrestrial Magnetism della Carnegie Institution di Washington ci siamo sforzati di trovare una risposta a queste domande misurando la velocità di rotazione di alcune galassie a varie distanze dal loro centro di rotazione. Da tempo si sa che fuori dal nucleo, particolarmente brillante, la luminosità di una tipica galassia a spirale cala rapidamente allontanandosi dal centro; se questa luminosità fosse un indice valido della presenza di massa la maggior parte di questa sarebbe concentrata verso il centro. Secondo la legge di Keplero che vale per le velocità orbitali dei corpi del sistema solare, al di fuori del nucleo questa velocità rotazionale dovrebbe diminuire in proporzione inversa alla radice quadrata della distanza. Si è trovato invece che le velocità orbitali, in un campione rappresentativo di galassie a spirale, si mantengono costanti al crescere della distanza dal centro, e a volte aumentano leggermente, almeno fin dove si riescono ancora a effettuare le misurazioni. Questo risultato inatteso indica che la diminuzione della massa luminosa con la distanza dal centro è compensata da un aumento della massa non luminosa.

I nostri risultati, uniti a quelli di numerosi altri ricercatori che hanno affrontato il problema della massa da punti di vista diversi, permettono oggi di affermare con una certa sicurezza che la distribuzione di luminosità non è un indice valido della distribuzione di massa, né all'interno delle galassie, né per quanto riguarda l'universo nel suo complesso. È evidente che qualcosa come il 90 per cento della massa dell'universo non irradia su nessuna lunghezza d'onda con intensità sufficiente a essere rilevata dalla Terra. Inizialmente gli astronomi consideravano la componente non luminosa dell'universo come «materia mancante», ma oggi è chiaro che questa massa, lungi dal mancare, è semplicemente invisibile. Questa materia oscura po-

Le galassie a spirale, di cui l'autore ha cercato di determinare la massa invisibile, si dividono in tre classi fondamentali: Sa, Sb e Sc. All'interno di ciascuna classe sia le dimensioni sia la luminosità delle singole galassie variano in misura notevole. In queste fotografie in negativo, in ognuna della quali 1,2 centimetri equivalgono a 10 chiloparsec (32 600 anni luce), compaiono nove galassie. Le tre Sa sono in alto, le Sb al centro e le Sc in basso. Nella successione il nucleo diventa via via più piccolo rispetto al disco e la struttura a spirale sempre più pronunciata.

Per le velocità orbitali dei pianeti del sistema solare, in cui più del 99 per cento della massa è concentrato nel Sole, la legge di Keplero fornisce questa curva. La velocità orbitale è inversamente proporzionale alla radice quadrata di *r*, la distanza media di un pianeta dal Sole, qui espressa in unità astronomiche (una U.A. è pari alla distanza media tra la Terra e il Sole). Plutone, che si trova a 39 U.A. dal Sole, ne dista 100 volte più di Mercurio, a 0,39 U.A., la cui velocità orbitale è di 47,9 chilometri al secondo. Per questo la velocità di Plutone sarà 10 volte più piccola: 4,7 chilometri al secondo (47,9 × 1/$\sqrt{100}$). I risultati dell'autore indicano che le velocità orbitali delle stelle di una galassia a spirale deviano fortemente da una distribuzione kepleriana.

trebbe esistere sotto forma di stelle molto deboli di piccola massa, di pianeti di grandi dimensioni, come Giove, o di buchi neri sia di grande sia di piccola massa. Tra gli altri possibili candidati si possono citare i neutrini (ammesso, come suggeriscono alcune ricerche recenti, che posseggano una massa), oppure particelle ipotetiche, come i monopoli magnetici e i gravitini.

All'inizio del secolo per tutta la comunità degli astronomi era ragionevole supporre che la distribuzione di materia luminosa corrispondesse dovunque a quella di massa. Circa 50 anni fa, però, Sinclair Smith e Fritz Zwicky del California Institute of Technology scoprirono che in alcuni grandi ammassi di galassie i singoli membri si muovono così velocemente che l'attrazione gravitazionale reciproca non basta a evitare la dispersione dell'ammasso. Esistono allora solo due possibilità: o questi ammassi dovrebbero disgregarsi oppure deve esistere una quantità di materia oscura sufficiente per tenerli assieme. Quasi tutti gli indizi disponibili tendono a far pensare che gli ammassi di galassie siano configurazioni stabili e quindi le osservazioni di Smith e Zwicky hanno introdotto la prima prova che questi ammassi contengono sia materia luminosa sia materia non luminosa.

Le ricerche compiute recentemente da numerosi astronomi hanno rafforzato questa conclusione. Gli studi condotti sia sulla dinamica di galassie singole, tra cui la nostra, sia su coppie, gruppi e ammassi di galassie evidenziano tutti la presenza di una inosservabile, ma onnipresente componente della massa totale, che è stata rivelata unicamente sulla base dei suoi effetti gravitazionali.

Da diversi anni W. Kent Ford, Jr., Norbert Thonnard, David Burstein e io stiamo cercando di scoprire nuovi dati sulla distribuzione della massa nell'universo studiando la distribuzione della materia nelle galassie di struttura simile a quella della nostra, nella classe cioè delle galassie a spirale. Abbiamo scelto questa via perché la geometria delle galassie a spirale si presta particolarmente bene all'identificazione della massa, sia luminosa sia non luminosa, e anche perché i grandi telescopi moderni, dotati di spettrografi a tubi convertitori d'immagine, permettono l'osservazione completa di una singola galassia con un'esposizione di circa tre ore. Ma prima di passare a descrivere le nostre osservazioni potrà essere utile soffermarsi su come i corpi celesti rispondano alla forza gravitazionale che agisce su di loro e come questa risposta permetta di comprendere la distribuzione su grande scala della materia.

Verso la fine del XVII secolo Robert Hooke sospettò che i pianeti fossero soggetti a una forza gravitazionale proveniente dal Sole di intensità inversamente proporzionale al quadrato della distanza. In seguito Isaac Newton si rese conto che tra una coppia qualsiasi di oggetti nell'universo ha luogo una mutua attrazione proporzionale al prodotto delle loro masse e inversamente proporzionale al quadrato della distanza reciproca. In altri termini raddoppiando per esempio la distanza tra i due oggetti, l'attrazione reciproca si riduce a un quarto.

Per quanto riguarda i pianeti che orbitano attorno al Sole, in cui è concentrata praticamente tutta la massa del sistema, alla diminuzione dell'attrazione gravitazionale con la distanza corrisponde una diminuzione della velocità necessaria per mantenere il pianeta sulla sua orbita. Per questo Mercurio a 0,39 unità astronomiche dal Sole (cioè a 0,39 volte la distanza media della Terra dal Sole) ha una velocità orbitale di circa 47,9 chilometri al secondo, mentre quella di Plutone, che si trova 100 volte più lontano, a una distanza media di 39,5 unità astronomiche, è pari solo a un decimo, cioè a 4,7 chilometri al secondo. Le galassie a spirale, dal canto loro, ruotano su se stesse perché conservano ancora il momento angolare e il momento angolare orbitale delle nubi di gas primordiali da cui si sono formate.

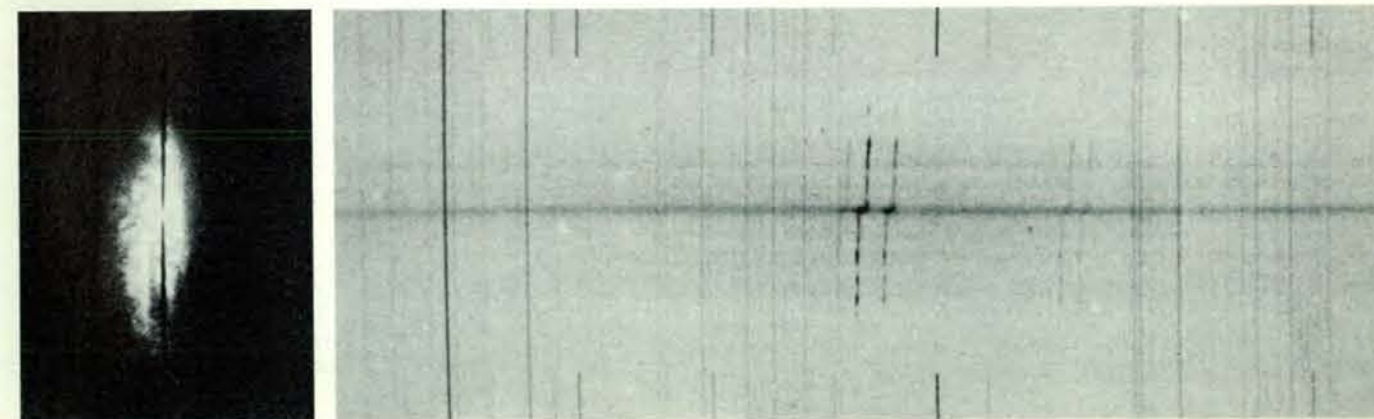
In una galassia a spirale gas, polvere e stelle che si trovano nel disco della galassia (e tutti i pianeti e i satelliti associati) si trovano in orbita attorno a un centro comune. Gas e stelle, come i pianeti nel sistema solare, si muovono in risposta all'attrazione gravitazionale combinata di tutte le altre masse presenti. Se immaginiamo che la galassia sia uno sferoide, l'attrazione gravitazionale dovuta alla massa M_r compresa tra il centro e un oggetto di massa m in orbita equatoriale a una distanza r dal centro stesso è data dalla legge di Newton GmM_r/r^2, in cui G è la costante di gravitazione universale. Se la galassia non si espande né si contrae la forza gravitazionale deve essere esattamente uguale alla forza centrifuga che agisce sulla massa m a distanza r: $GmM_r/r^2 = mV_r^2/r$, dove V_r indica la velocità orbitale.

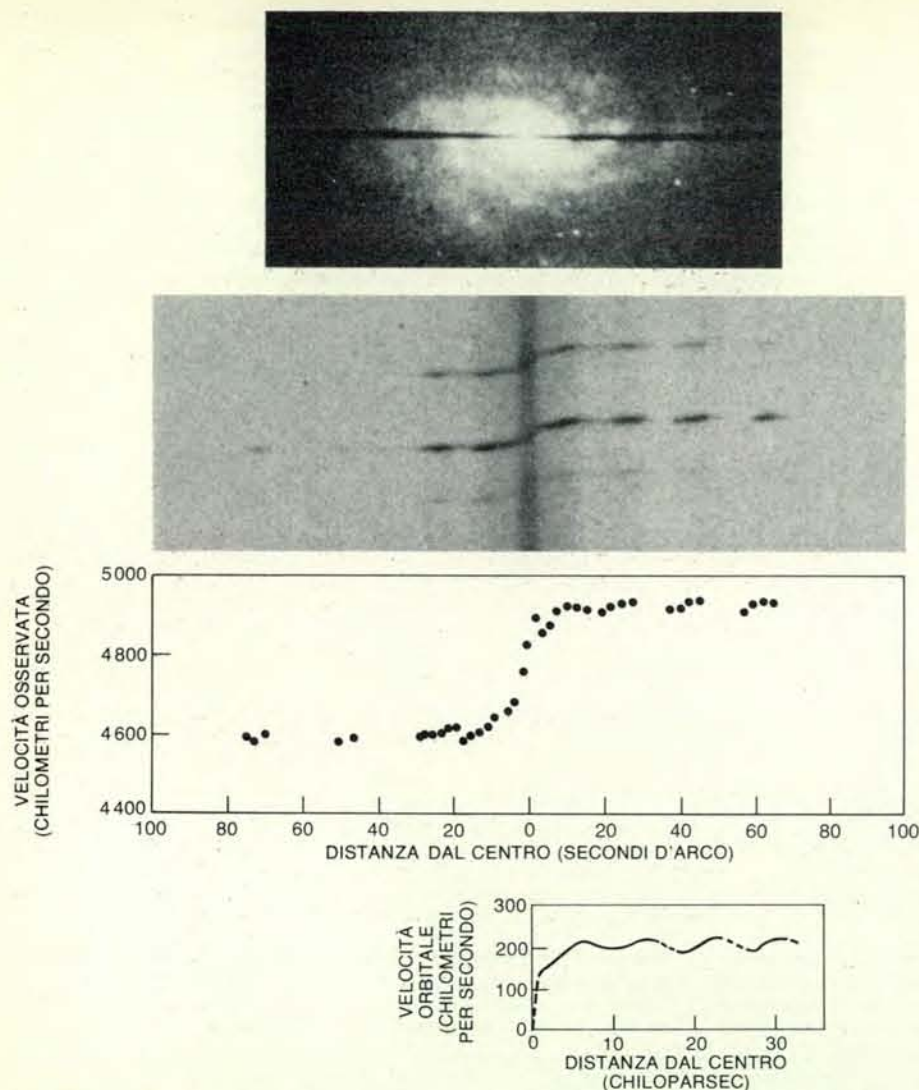
Ricavando V_r da questa equazione si può semplificare il valore di m e la velocità di un corpo situato a distanza r dal centro è determinata soltanto dalla massa M_r, contenuta nella sfera di raggio r. Se, come accade nel sistema solare, tutta la massa è praticamente raccolta vicino al centro le velocità orbitali diminuiscono, andando verso l'esterno, in proporzione inversa a r^2. Questa diminuzione della velocità orbitale è chiamata kepleriana da Giovanni Keplero, che fu il primo a enunciare le leggi del moto dei pianeti.

La luminosità di una galassia presenta un massimo pronunciato al centro e diminuisce rapidamente con la distanza radiale; per molto tempo gli astronomi hanno supposto che anche la massa si comportasse nello stesso modo. Per questo si prevedeva che le stelle a distanze crescenti dal centro avessero velocità orbitali kepleriane, ma fino a poco tempo fa erano poche le osservazioni a favore o contro questa tesi compiute nelle deboli regioni esterne delle galassie.

Anche se le galassie a spirale possono assumere un'incredibile varietà di forme, gli astronomi sono in grado di raggrupparle in tre classi secondo una classificazione molto utile proposta circa 60 anni fa da Edwin P. Hubble. In questo schema le galassie denominate Sa si distinguono per una grossa sporgenza centrale circondata da bracci molto regolari, strettamente arrotolati, in cui è difficile individuare «nodi», cioè regioni di particolare luminosità. Nelle galassie Sb il rigonfiamento centrale è meno pronunciato, i bracci sono più aperti e i nodi più evidenti, mentre le galassie Sc hanno un rigonfiamento centrale molto piccolo e bracci ben separati gli uni dagli altri, chiazzati da tratti luminosi molto marcati. Passando dalle galassie Sa alle galassie Sc si ha una diminuizione progressiva del rigonfiamento centrale, mentre il disco che gli ruota attorno diviene via via preponderante. Il fatto che il disco sia davvero in rotazione si deduce da considerazioni di dinamica molto semplici.

All'interno di ciascun tipo di galassia si possono poi riscontrare variazioni sistematiche di dimensioni e di luminosità; le galassie Sc, per esempio, variano da piccoli oggetti, di massa e luminosità molto ridotte, a galassie di enorme luminosità e massa. Per amore di completezza, quindi, lo studio della dinamica delle galassie a spirale dovrebbe prendere in considerazione non solo oggetti appartenenti a categorie morfologiche diverse, ma anche oggetti di diversa luminosità.

Solo per le stelle più vicine nella nostra stessa galassia è possibile rilevare il moto osservando la variazione di posizione della stella rispetto a elementi «fissi» della sfera celeste come stelle e galassie lontane. Anche solo nella grande galassia di Andromeda, la galassia a spirale più vicina a noi, una stella che si muovesse con una velocità orbitale di 200 chilometri al secondo (una velocità paragonabile a quella del Sole) impiegherebbe circa 20 000 anni per percorrere nel cielo un

Lo spettro della galassia a spirale NGC 7541 *(a destra)* è stato rilevato dall'autore e da W. Kent Ford, Jr., con il telescopio da quattro metri del Kitt Peak National Observatory. È una galassia a spirale di tipo Sc, distante 60 megaparsec. (Un megaparsec è pari a 3,26 milioni di anni luce.) Il tempo di esposizione è stato di 114 minuti. A sinistra la galassia come appare sullo schermo televisivo della sala comandi del telescopio. La riga scura che attraversa la galassia segue l'orientamento della fenditura dello spettrografo, che raccoglie campioni di luce.

Le righe di emissione dello spettrogramma di NGC 7541 provengono da due sorgenti: il cielo notturno e gli atomi delle nubi di gas che circondano le stelle luminose della galassia. Quasi tutte le righe dovute al cielo notturno, presenti su tutta l'estensione dello spettrogramma, sono dovute a molecole con ossidrili (OH) presenti nell'atmosfera terrestre. Alcune altre invece sono dovute a atomi di idrogeno e di ossigeno atmosferici. La rotazione di NGC 7541 sposta le righe di emissione del disco verso lunghezze d'onda maggiori (più rosse) o minori (più blu) a seconda che le stelle e i gas del disco, ruotando, si avvicinino o si allontanino dall'osservatore. Dato poi che tutta la galassia si allontana per l'espansione complessiva dell'universo, la riga alfa dell'idrogeno dovuta al gas galattico è spostata verso il rosso rispetto a quella dell'idrogeno nell'atmosfera; l'entità dello spostamento permette di valutare la velocità di recessione della galassia. L'inclinazione delle righe di emissione galattica indica che la velocità orbitale delle stelle e del gas del disco aumenta in proporzione alla distanza dal centro galattico.

Per misurare la rotazione di NGC 2998, una galassia Sc a 96 megaparsec di distanza nella costellazione dell'Orsa Maggiore, bisogna innanzitutto farne uno spettrogramma. Nell'illustrazione in alto si vede la galassia con sovrapposta la fenditura dello spettrografo come appare sullo schermo televisivo del telescopio da quattro metri di Kitt Peak. Più in basso compare la regione corrispondente alla riga alfa dell'idrogeno in uno spettrogramma ottenuto con un'esposizione di 200 minuti. I punti del grafico rappresentano le velocità attraverso il disco galattico misurate a partire dalla riga alfa dell'idrogeno. La galassia nel suo complesso si allontana a 4800 chilometri al secondo, ma il suo lato sinistro si avvicina a noi, mentre il destro si allontana. Infine si traccia una curva di rotazione per la galassia, unendo con una linea continua i valori di velocità sui due lati del disco e convertendo le distanze angolari sulla sfera celeste in distanze lineari nella galassia.

secondo d'arco, che è la minima separazione angolare rilevabile con mezzi ottici dalla Terra. Per studiare i moti all'interno delle galassie occorre quindi un sistema diverso, un sistema che sfrutta il fenomeno dello spostamento Doppler.

L'effetto Doppler consiste in uno spostamento nella frequenza delle onde emesse da una certa sorgente dovuto al suo moto di avvicinamento o di allontanamento dall'osservatore. Registrando lo spettro del nucleo brillante di una galassia a spirale le righe di assorbimento dovute alle stelle che ne fanno parte vengono spostate verso l'estremità rossa dello spettro rispetto alle stesse righe ottenute in un laboratorio terrestre. Questi spostamenti verso il rosso, visibili negli spettri di tutte le galassie, se si eccettua il caso di alcune galassie vicinissime, e osservati intorno al 1915 da V. M. Slipher del Lowell Observatory, sono la prima prova del fatto che l'universo sta espandendosi, trascinando con sé quasi tutte le galassie isolate che si allontanano così tra di loro e dalla nostra. In seguito alle ricerche di Smith e di Zwicky oggi si sa che in tutti i raggruppamenti, le coppie, gli ammassi di galassie, il campo gravitazionale locale prevale sulla tendenza generale all'espansione, così che questi agglomerati di materia relativamente densa rimangono legati: la distanza tra gli ammassi di galassie aumenta ma quella tra le singole galassie di uno stesso ammasso rimane approssimativamente costante. Slipher, inoltre, si rese conto del fatto che gli spettri di galassie singole sono in grado di fornire anche informazioni sul moto delle stelle e dei gas all'interno della galassia.

Se il disco di una galassia a spirale è orientato in modo che il piano che lo contiene sia fortemente inclinato rispetto alla nostra direzione di osservazione, la rotazione della galassia trascinerà le stelle e il gas che si trovano da un lato del nucleo galattico verso la nostra galassia e le stelle e il gas sull'altro lato in direzione opposta. Le righe spettrali della materia in avvicinamento risulteranno quindi spostate verso il blu, cioè verso le frequenze più alte, mentre le righe relative alla materia che si allontana subiranno uno spostamento verso il rosso, cioè una diminuizione di frequenza. Rilevando la frequenza di una certa riga in vari punti del disco è perciò possibile misurare, oltre alla distanza angolare del punto considerato dal nucleo galattico, la componente lungo la linea di vista della velocità della materia a quella distanza.

È difficile effettuare per via spettroscopica misurazioni delle velocità di stelle singole, le quali sono deboli anche in galassie relativamente vicine alla nostra. Per questo motivo non osserviamo le stelle, ma la luce proveniente da certe nubi di gas, ricche di idrogeno ed elio, che circondano alcune stelle calde. Lo spettro di queste nubi è costituito da righe di emissione luminose, emesse da un atomo eccitato quando un elettrone cade da un livello energetico più alto a uno più basso. Oltre alle righe di emissione dell'idrogeno e dell'elio compaiono di solito altre righe molto evidenti dovute ad atomi di azoto e di zolfo ionizzati una volta, cioè privati di un elettrone. Queste righe si chiamano «proibite» perché hanno origine da atomi presenti nel gas estremamente rarefatto dello spazio; nei laboratori sulla Terra questi atomi ionizzati una volta abbandonano rapidamente lo stato eccitato collidendo con altri atomi prima che la transizione proibita abbia luogo.

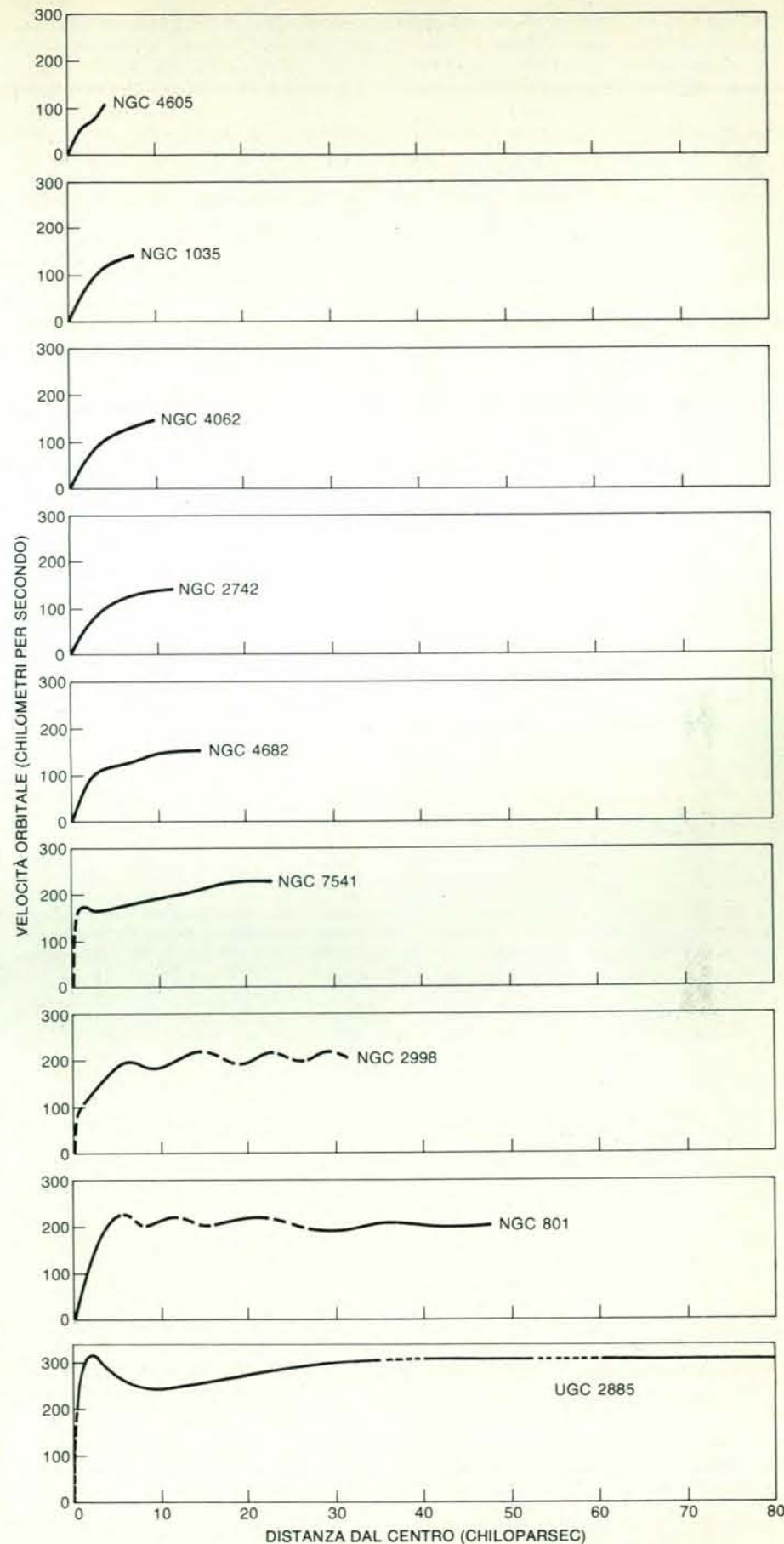
Fino a poco tempo fa era impossibile ottenere spettrogrammi a elevata risoluzione delle regioni esterne, e quindi deboli, delle galassie; è solo grazie al fatto che oggi sono disponibili grandi telescopi ottici ad alta risoluzione spettrografica a fenditura lunga e validi dispositivi elettronici per la formazione di immagini che il nostro programma di osservazione è diventato attuabile. Sei anni fa i miei colleghi e io abbiamo iniziato a misurare le velocità di rotazione trasversalmente al disco luminoso di galassie a spirale inclinate in maniera opportuna; il nostro obiettivo era quello di studiare la dinamica interna e la distribuzione di massa in galassie singole in funzione della loro morfologia. Fino a oggi siamo arrivati a osservare 60 galassie a spirale, 20 per ciascuna delle classi principali: Sa, Sb e Sc. Abbiamo scelto galassie che appartenessero a un tipo ben definito, che fossero opportunamente inclinate sul piano del cielo, in modo che una grande componente della velocità orbitale venisse a trovarsi lungo la linea di vista, che avessero un diametro angolare non superiore alla lunghezza della fenditura dello spettrografo e che ricoprissero un'ampia gamma di luminosità per ogni tipo.

La maggior parte degli spettri sono stati ottenuti con telescopi da quattro metri, quello del Kitt Peak National Observatory in Arizona e quello del Cerro Tololo Inter-American Observatory in Cile. Alcuni spettri sono stati registrati anche con il telescopio di 2,5 metri del Las Campanas Observatory, sempre in Cile.

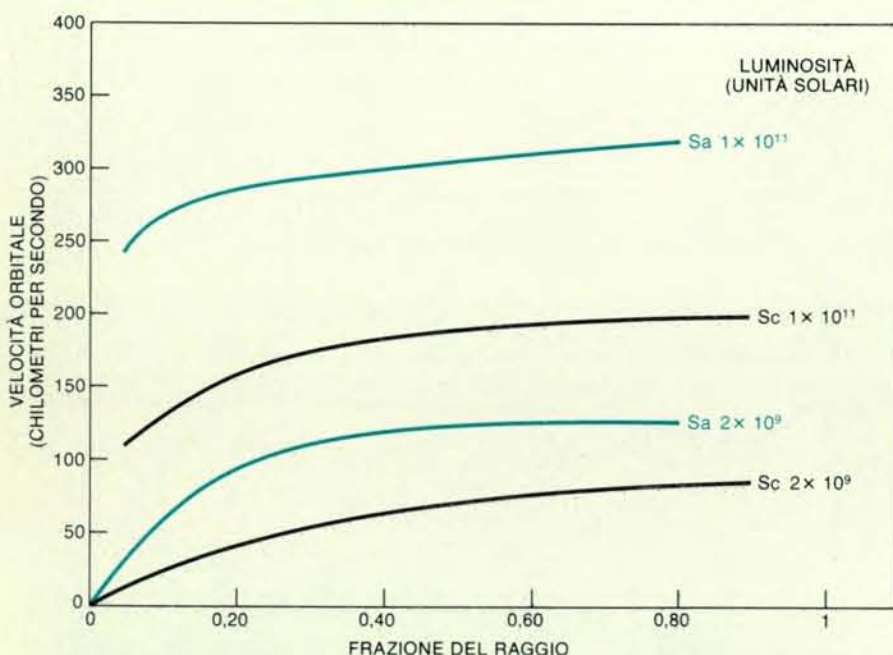
I fotoni provenienti dalla sorgente galattica, dopo essere passati attraverso la fenditura ed essere stati dispersi da un reticolo di diffrazione, vengono concentrati su un tubo convertitore di immagini «Carnegie» (RCA C33063), che li moltiplica per 10 o più prima che vengano registrati dall'emulsione fotografica. Per le esposizioni, della durata di due o tre ore, si usano lastre Kodak IIIa-J. Per portare la loro sensibilità all'altezza di quella del tubo convertitore di immagine, le lastre sono state sottoposte a una «cottura» di due ore a 65 gradi centigradi in un'atmosfera di azoto addizionato con il 2 per cento di idrogeno, e a una preesposizione a lampi di luce. In mancanza del tubo convertitore e di questi metodi di sensibilizzazione i tempi di esposizione sarebbero risultati assolutamente proibitivi: dalle 20 alle 60 ore.

In genere per ogni galassia vengono esposte due lastre: in una si fa coincidere la fenditura dello spettrografo con l'asse maggiore della galassia, ottenendo uno spettrogramma in cui a ogni punto corrisponde una certa regione del disco galattico. In questo caso nell'immagine sviluppata risultano evidenti gli spostamenti Doppler delle righe di emissione. Nella seconda esposizione la fenditura dello spettrografo viene allineata lungo l'asse minore del disco galattico. In questo caso non si rileva alcuno spostamento Doppler, perché le velocità orbitali nelle regioni osservate sono perpendicolari alla linea di vista. È proprio l'assenza di qualsiasi spostamento delle righe a garantire che i moti in esame sono davvero orbitali.

Abitualmente, per avere una scala di riferimento per misurare lo spostamento delle righe di emissione negli spettri galattici, gli astronomi fotografavano lungo i margini dello spettro le righe dovute a una lampada al neo. Noi, invece, siamo riusciti a evitare il ricorso a questo procedimento valutando gli spostamenti direttamente sulla base delle righe non spostate e presenti su ogni lastra delle molecole con gruppi ossidrilici (OH) presenti nell'atmosfera terrestre. Molti astronomi hanno adottato dispositivi, particolarmente sofisticati, di scansione delle lastre, soprattutto per misurare la posizione delle righe, in particolare quelle più deboli. Noi, invece, misuriamo ancora la posizione delle righe di emissione con l'ausilio di un microscopio il cui tavolino si può muovere in due direzioni; con questo sistema siamo in grado di rilevare posizioni su ciascuna coordinata con la precisione di un micrometro, un valore estremamente soddisfacente.

Nel nostro lavoro definiamo «raggio nominale» di una galassia la distanza dal centro alla quale la sua luminosità superficiale cade al di sotto della soglia di rile-

Queste curve di rotazione indicano la variazione delle velocità orbitali al crescere della distanza dal centro in nove galassie Sc di luminosità crescente dall'alto verso il basso. Via via che la luminosità aumenta le galassie diventano sempre più grandi, e così pure le velocità orbitali, mentre i gradienti di velocità in prossimità del centro galattico si fanno progressivamente più ripidi.

vabilità su lastre ottenute con il telescopio Schmidt da 48 pollici di Monte Palomar, un valore pari alla luminosità di una sorgente della 25ª magnitudine che copra un'area di un secondo d'arco di lato. Per determinare la distanza, e quindi le dimensioni effettive degli oggetti in esame, adottiamo per la costante di Hubble (che rappresenta il tasso di espansione dell'universo) un valore di 50 chilometri al secondo per megaparsec. (Un megaparsec è pari a 3,26 milioni di anni luce.)

A partire dalle velocità misurate delle righe di emissione più intense calcoliamo una curva di rotazione regolare, facendo la media tra i valori delle velocità di avvicinamento e di allontanamento relativi ai due lati della galassia. Anche se ogni galassia ha il proprio diagramma di velocità diverso da quello di tutte le altre, le tendenze sistematiche che emergono sono impressionanti. Al crescere della luminosità, in primo luogo, le galassie sono più grandi, le velocità orbitali più alte e il gradiente della velocità lungo il nucleo galattico è più ripido. In ogni tipo di galassia, poi, la rotazione si svolge secondo uno schema che rivela alcune proprietà caratteristiche. Nelle galassie Sa più luminose, ad esempio, la velocità di rotazione a metà del raggio galattico è superiore di oltre il 50 per cento di quella delle galassie Sc di pari luminosità, e all'interno di questa classe le galassie più luminose ruotano, a distanze equivalenti dal centro, a velocità quasi doppia di quelle più deboli, che hanno una luminosità pari solo a un centesimo.

Dalle nostre osservazioni si fa strada una conclusione ben precisa. Tutte o quasi le curve di rotazione ricavate sono piatte oppure tendono a salire verso valori di velocità sempre maggiori fino ai limiti di visibilità delle galassie: non esiste nessuna regione, se non di dimensioni trascurabili, in cui le velocità di rotazione diminuiscano con la distanza dal centro, come sarebbe da prevedere se la massa fosse concentrata verso il centro. La conclusione è inevitabile: contrariamente a quanto si può affermare per la luminosità, la massa non è raccolta in vicinanza del centro delle galassie a spirale. La distribuzione della luminosità all'interno di una galassia, quindi, non è affatto indicativa della distribuzione della massa.

Sulla base delle velocità di rotazione si può calcolare che le masse delle galassie che abbiamo preso in esame siano comprese tra le 6×10^9 e le 2×10^{12} masse solari limitatamente alla regione compresa in un raggio pari a quello del disco visibile. A tutt'oggi non è ancora possibile precisare il valore della massa totale di nessuna di queste galassie, perché non siamo in grado di determinare nessun tipo di limitazione per questo valore. Al contrario, la massa compresa entro una distanza radiale determinata cresce linearmente all'aumentare della distanza e, contro le aspettative più ovvie, non dà segno di convergere a un qualsiasi limite avvicinandosi al bordo del disco visibile. La proporzionalità diretta tra raggio e massa indica che ogni successivo guscio sferico di materia deve contenere la stessa massa di ogni altro guscio dello stesso spessore. Dato perciò che il volume di ogni guscio successivo aumenta con il quadrato del raggio, la densità della materia nei gusci successivi deve risultare inversamente proporzionale al quadrato del raggio, perché rimanga costante il prodotto tra densità e volume.

Il modello teorico meno contrastante con le idee oggi generalmente accettate riguardo alle galassie spiega le curve di rotazione osservate inglobando ogni galassia a spirale in un «alone» di materia sferico, che si estende ben al di là del disco galattico visibile: è l'attrazione gravitazionale dovuta a questa massa invisibile che impedisce alla velocità orbitale di diminuire al crescere della distanza dal centro galattico. Può sembrare fastidioso, magari, che le osservazioni non forniscano praticamente nessun elemento utile per descrivere con una certa precisione la distribuzione di questa materia oscura invisibile. È possibile però affermare che la massa invisibile non fa parte del fondo di materia che riempie tutto l'universo, perché non è distribuita in modo uniforme, ma è fortemente concentrata attorno alle galassie. La densità dell'alone, infatti, decresce, anche se lentamente, allontanandosi dal centro galattico, mantenendosi però da 100 a 1000 volte più alta della densità media dell'universo anche a grandi distanze radiali.

Le velocità orbitali di galassie di tipo Sc di diversa luminosità in funzione del loro raggio otticamente visibile. Le luminosità, espresse in unità di luminosità solare, variano di due ordini di grandezza mentre le velocità orbitali, a parità di distanza dal centro, aumentano con la luminosità.

Un confronto tra galassie Sa e galassie Sc rivela che, a parità di luminosità, le velocità orbitali sono decisamente maggiori nelle Sa a tutte le distanze dal centro. Ciò significa che le galassie Sa contengono una maggiore quantità di massa per unità di luminosità rispetto alle galassie Sc.

Esistono altri modelli che cercano di spiegare le elevate velocità orbitali, ma sono tutti meno soddisfacenti di quello che prevede un singolo alone di materia non luminosa. Se tutta la materia oscura necessaria fosse concentrata in un disco, questo diventerebbe rapidamente instabile e assumerebbe una forma allungata. L'idea che gli aloni sono necessari per rendere stabile un disco è stata proposta da Jeremiah P. Ostriker e P. J. E. Peebles della Princeton University.

Modelli di galassie a spirale nei quali la massa è ripartita in un nucleo, un rigonfiamento attorno a questo nucleo, un disco e un alone, sono in grado di riprodurre gli effetti dinamici osservabili. John N. Bahcall e Raymond M. Soneira dell'Institute for Advanced Study, Maarten Schmidt del Cal Tech e S. Casertano della Scuola normale superiore di Pisa hanno elaborato modelli particolarmente interessanti. Ma l'idea più spregiudicata per spiegare valori così elevati delle velocità di rotazione è forse quella proposta indipendentemente da Joel E. Tohline della Louisiana State University e da M. Milgrom e J. Bekenstein del Weizmann Institute of Science, che hanno ipotizzato la necessità di modificare per le grandi distanze la teoria newtoniana della gravitazione, il che permetterebbe di spiegare il valore elevato delle velocità rotazionali nelle galassie a tali distanze anche in assenza di qualsiasi tipo di massa invisibile.

Un ulteriore indizio del fatto che nelle galassie a spirale la materia si muove con velocità angolari più elevate del previsto ci giunge dall'emissione radio a 21 centimetri proveniente dall'idrogeno neutro (non ionizzato) presente nel disco galattico. Fin dai primi studi di questa radiazione, condotti su un piccolo campione di galassie a spirale da Morton S. Roberts del National Radio Astronomy Observatory, è parso chiaro che le velocità di rotazione dell'idrogeno sono elevate. Oggi, utilizzando i radiotelescopi multipli, e in particolare la schiera di telescopi di Westerbork in Olanda e il Very Large Array di Socorro, nel New Mexico, è possibile raggiungere e perfino superare il potere risolutivo dei migliori telescopi ottici e, quindi, studiare la distribuzione dell'idrogeno in galassie analoghe a quelle che noi abbiamo osservato nella banda ottica dello spettro. A questo proposito Albert Bosma dell'Università statale di Leida ha messo in luce il fatto che, per una vasta gamma di tipi di galassie, le velocità orbitali dell'idrogeno neutro si mantengono elevate anche a notevoli distanze dal centro galattico.

In generale i diametri apparenti delle galassie risultano abbastanza simili nei rilevamenti ottici e in quelli radio, ma per una piccola serie di galassie l'idrogeno si spinge fino a distanze dal centro parecchie volte più grandi di quelle raggiunte dalle stelle visibili. In questi oggetti, quindi, è possibile determinare il potenziale gravitazionale oltre i limiti di visibilità della galassia con mezzi ottici. In diversi casi, però, l'idrogeno non rimane su un piano, ma viene fortemente deformato in prossimità del margine del disco visibile. Per questo motivo non è sicuro che le velocità del gas rilevate alle massime distanze dal centro siano effettivamente velocità orbitali circolari e non si debbano a moti più complessi.

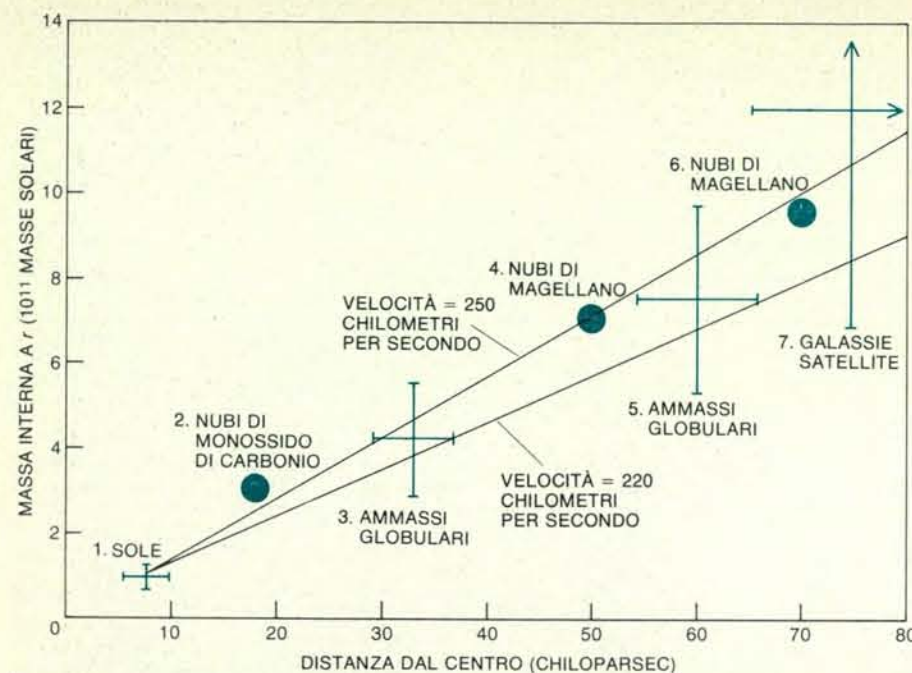
Renzo Sancisi dell'Università di Groninga, che si occupa da qualche tempo di galassie «deformate» di questo tipo, ha avanzato l'ipotesi che le velocità orbitali decrescano davvero, oltre i limiti della

$$\frac{GmM_r}{r^2} = \frac{mV_r^2}{r} \rightarrow M_r = \frac{rV_r^2}{G}$$

FORZA GRAVITAZIONALE — FORZA CENTRIFUGA

	NGC 1035		NGC 2998	
RAGGIO (CHILOPARSEC)	VELOCITÀ V_r (CHILOMETRI PER SECONDO)	MASSA INTERNA M_r (10^{10} MASSE SOLARI)	VELOCITÀ V_r (CHILOMETRI PER SECONDO)	MASSA INTERNA M_r (10^{10} MASSE SOLARI)
0,5	39	0,018	87	0,088
1	65	0,098	102	0,24
2	91	0,39	126	0,74
3	107	0,80	142	1,4
5	123	1,8	182	3,9
8	135	3,4	204	7,7
20			214	21
30			214	32

È possibile calcolare la massa all'interno di un dato raggio uguagliando l'attrazione gravitazionale e la forza centrifuga a ogni distanza r dal centro della galassia. Nelle equazioni G è la costante di gravitazione universale, m la massa che si trova a distanza r, M_r la massa a distanza minore di r e V_r la velocità orbitale di m; la massa nella sfera di raggio r aumenta linearmente con r. La tabella dà le masse comprese entro il raggio r per due galassie Sc: NGC 1035, poco luminosa, e NGC 2998, di luminosità elevata. Per ogni distanza dal centro la galassia più luminosa ha velocità orbitale superiore e deve, quindi, avere una massa maggiore compresa entro quella distanza.

Confrontando le masse comprese entro sfere di raggio crescente, sia per galassie molto luminose sia per quelle più deboli, si nota che la massa cresce in modo approssimativamente proporzionale alla distanza dal centro e non dà segno di convergere a un qualche limite avvicinandosi al margine della galassia rilevabile con mezzi ottici. Per ogni distanza dal centro le galassie di tipo Sa rivelano una massa e quindi una densità maggiore rispetto alle galassie Sc di pari luminosità.

Diversi studi condotti sulla nostra galassia permettono di formulare qualche valutazione della massa compresa in una distanza *r* misurata in chiloparsec dal centro galattico. Queste valutazioni derivano dal rilevamento delle distanze medie e delle velocità orbitali di molti corpi diversi. Il valore adottato per la velocità orbitale del Sole, a una distanza di otto chiloparsec, è di 220 chilometri al secondo. Per ottenere il secondo punto del grafico si è sfruttata la velocità media di alcune nubi di ossido di carbonio a una distanza media di 18 chiloparsec, misurata da Leo Elits dell'Università del Maryland a College Park. Il terzo e il quinto punto derivano dalle velocità di alcuni ammassi globulari appartenenti all'«alone» della nostra galassia e situati a due diverse distanze medie dal nucleo. Le velocità degli ammassi più vicini sono state analizzate da Carlos Frenk, che all'epoca lavorava presso l'Università di Cambridge, e da Simon White dell'Università della California a Berkeley. Lo studio delle velocità degli ammassi più lontani è opera di F. D. A. Hartwick dell'Università di Victoria e di Wallace L. W. Sargent del California Institute of Technology. Il quarto è stato ottenuto dalle valutazioni delle velocità delle Nubi di Magellano, le galassie più vicine alla nostra, compiute da Tadayki Murai e Mitsuaki Fujimoto dell'Università di Nagoya. Il sesto rappresenta le stime delle distanze e delle velocità delle Nubi di Magellano effettuate indipendentemente da D. N. C. Lin del Lick Observatory e da Donald Lynden-Bell dell'Università di Cambridge. L'ultimo punto è ricavato dalle velocità di altre galassie satellite più lontane, secondo i risultati degli studi di Jaan Einasto e dei suoi collaboratori dell'Accademia delle scienze dell'Estonia. L'estensione delle linee verticali indica l'arco dei valori per orbite di geometrie diverse. Queste misure suggeriscono che le velocità di rotazione nella nostra galassia sono comprese tra 220 e 250 chilometri al secondo e rimangono costanti fino a una distanza di 80 chiloparsec, e cioè fino a circa 10 volte la distanza che separa il Sole dal centro della Galassia. La massa compresa in questa regione è probabilmente circa 10 volte la massa contenuta nella sfera di raggio pari alla distanza del Sole dal centro, e cioè circa 10^{12} masse solari.

galassia visibile, ma probabilmente solo di poco, forse 20 chilometri al secondo, cioè di un 10 per cento circa, per poi rimanere costanti a distanze molto maggiori. Le osservazioni radio, comunque, proseguono, e dovrebbero fornire informazioni molto preziose sulle regioni molto esterne delle galassie.

È una fortuna, per gli studiosi della struttura e dei fenomeni galattici, poter indagare le proprietà di galassie molto lontane e poi ritornare alla propria per vedere se ha le stesse caratteristiche. Fino a non molto tempo fa, infatti, gli astronomi erano convinti che il Sole, a circa otto chiloparsec dal centro della Galassia, si trovasse ai margini e che la Galassia fosse di dimensioni piuttosto limitate, mentre oggi tutti gli indizi sono concordi nel suggerire che anche la nostra galassia si estenda ben oltre la posizione del Sole e che la sua massa continui a crescere.

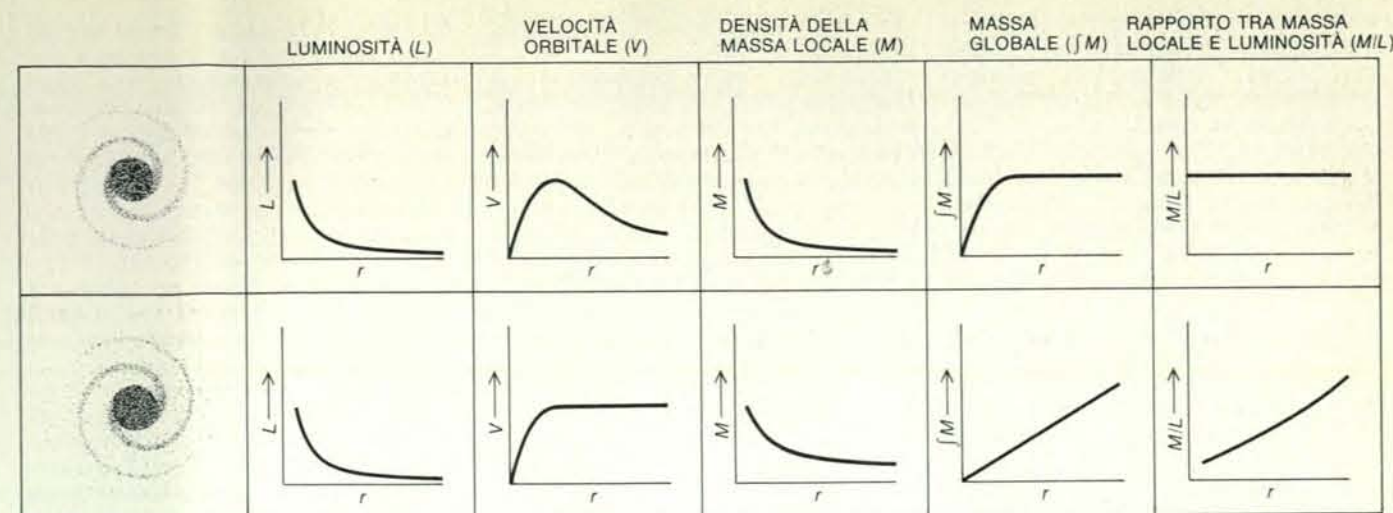
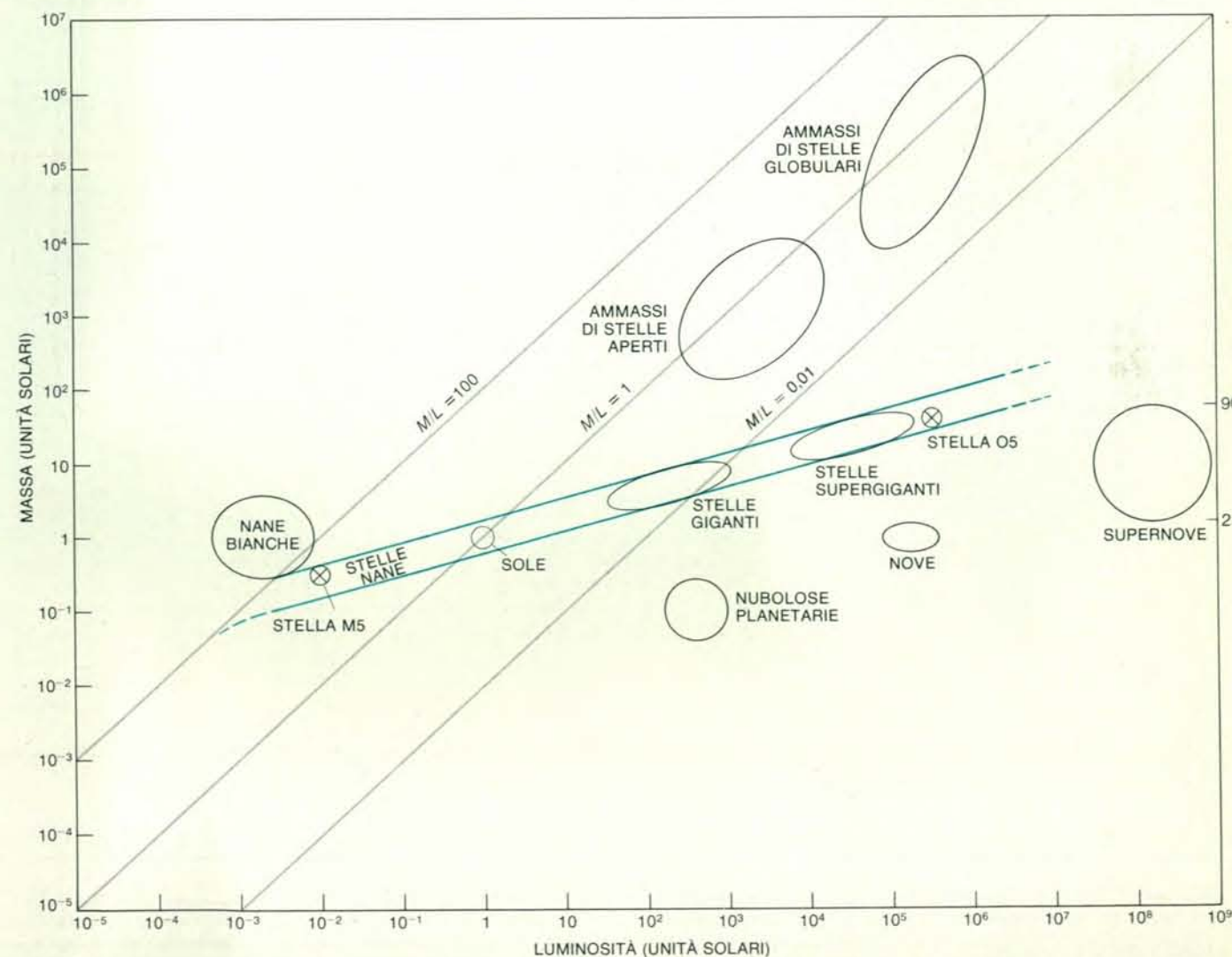
La velocità con cui il Sole percorre la sua orbita attorno al centro della Galassia è stimata, da James E. Gunn e Gillian R. Knapp di Princeton e da Scott D. Tremaine del Massachusetts Institute of Technology, intorno ai 220 chilometri al secondo circa, mentre altre valutazioni arrivano fino ai 260 chilometri al secondo. Utilizzando il valore più basso si ricava per la massa compresa tra il Sole e il centro della Galassia un valore di circa 10^{11} masse solari. Poiché si sa che a distanze dal centro galattico maggiori di quella del Sole si trova una massa abbastanza notevole di materia, la massa galattica entro un raggio di 100 chiloparsec potrebbe arrivare anche a 10^{12} masse solari, il che permetterebbe di annoverare la nostra galassia tra le più grandi del suo tipo.

Circa 30 anni fa Jan H. Oort dell'Osservatorio di Leida dimostrò che la massa osservabile delle stelle e del gas presenti nel disco galattico nelle vicinanze del Sole è poco più di metà di quella necessaria per spiegare l'attrazione gravitazionale del disco sulle stelle più lontane dal piano galattico centrale. Questo studio fornì la prima prova del fatto che anche la nostra galassia contiene una certa quantità di massa che non è luminosa.

Prove, più recenti, derivano dalle velocità orbitali di oggetti che si trovano nel piano galattico molto più lontani del Sole rispetto al centro. Si tratta di misurazioni difficili, ma in qualche caso fortunato si è riusciti a ricavare valori attendibili. Leo Blitz dell'Università del Maryland a College Park, per esempio, ha determinato le velocità di alcune nubi di monossido di carbonio che distano circa 16 chiloparsec dal centro galattico. Queste velocità, combinate a quelle di certe nubi di idrogeno valutate da Blitz stesso e da Shrinivas Kulkarni e Carl E. Heiles dell'Università della California a Berkeley, forniscono una curva di rotazione che continua a salire con la distanza dal centro galattico.

Per valutare la massa presente a distanze ancora maggiori F. D. A. Hartwick dell'Università di Victoria, Wallace L. W. Sargent del Cal Tech, Carlos Frenk dell'Università di Cambridge e Simon White dell'Università della California a Berkeley hanno misurato le velocità di alcuni ammassi globulari che si trovano nell'alone della nostra galassia, e precisamente di un campione di ammassi a 30 chiloparsec e di uno a 60 chiloparsec dal centro. Le loro ricerche rivelano che la massa della Galassia continua a crescere in modo all'incirca proporzionale al raggio fino alla distanza media degli ammassi.

Con molta fatica e un pizzico di fantasia si riesce a campionare il potenziale gravitazionale a distanze ancora maggiori. La nostra galassia, infatti, non è sola nello spazio; la accompagna un corteo di piccole galassie satelliti. Le orbite delle due galassie satelliti più vicine, la Grande e la Piccola Nube di Magellano, distanti dal centro della Galassia un po' meno di 60 chiloparsec, sono note in maniera molto approssimativa, ma Tadayuki Murai e Mitsuaki Fujimoto dell'Università di Nagoya, D. N. C. Lin del Lick Observatory e Donald Lynden-Bell dell'Università di Cambridge hanno proposto alcuni modelli orbitali, dai quali si deducono valori di massa in accordo con quelli ricavati dagli ammassi globulari.

Per distanze ancora maggiori Jaan Einasto e colleghi dell'Accademia delle Scienze dell'Estonia si sono basati su un insieme di ammassi globulari particolarmente lontani e di galassie satelliti per ricavare i valori di massa fino a oltre 80 chiloparsec di distanza. Combinando i risultati di queste analisi si ottiene un quadro complessivo della Galassia in cui le velocità orbitali rimangono comprese tra i 220 e i 250 chilometri al secondo fino a una distanza dal centro galattico pari a quasi dieci volte quella del Sole. Una distribuzione di massa di questo genere, inevitabile se si vuole che la nostra galassia non si discosti troppo da

Tutte le proprietà delle galassie reali, tranne la luminosità, differiscono notevolmente da quelle delle galassie ipotetiche. La tipica galassia a spirale reale, in basso, è dotata di un alone oscuro di massa notevole, mentre quella ipotetica, in alto, è completamente priva di alone: in base alle previsioni, la sua luminosità superficiale diminuisce rapidamente; le velocità orbitali al di fuori del nucleo diminuiscono in forma kepleriana; la densità di massa locale scende proporzionalmente alla luminosità; la massa integrale raggiunge un valore limite e il rapporto tra massa e luminosità rimane approssimativamente costante all'aumentare della distanza radiale. In una galassia reale la presenza di un alone oscuro modifica radicalmente queste proprietà, tranne l'aspetto ottico della galassia. Le velocità orbitali restano elevate, la densità di massa locale decresce solo molto lentamente, mentre la massa complessiva aumenta linearmente al crescere del raggio e il rapporto tra massa e luminosità cresce costantemente man mano che l'alone contribuisce di più alla massa e la luminosità del disco scende alla soglia di rilevabilità.

Sono messe a confronto la massa e la luminosità di alcuni componenti tipici di una galassia a spirale come la nostra. La massa e la luminosità del Sole fungono da unità. Espresso in queste unità il valore del rapporto tra massa e luminosità *M/L* per le stelle normali va da 30 circa, per stelle nane, vecchie e fredde (di tipo M5), fino a circa 10^{-4} per le stelle giovani e calde (del tipo O5). Solo le nane bianche, molto dense, hanno un valore di *M/L* più alto di 100. Per l'alone di una galassia, dove *M/L* è nell'ordine delle centinaia, sono necessarie altre classi di oggetti.

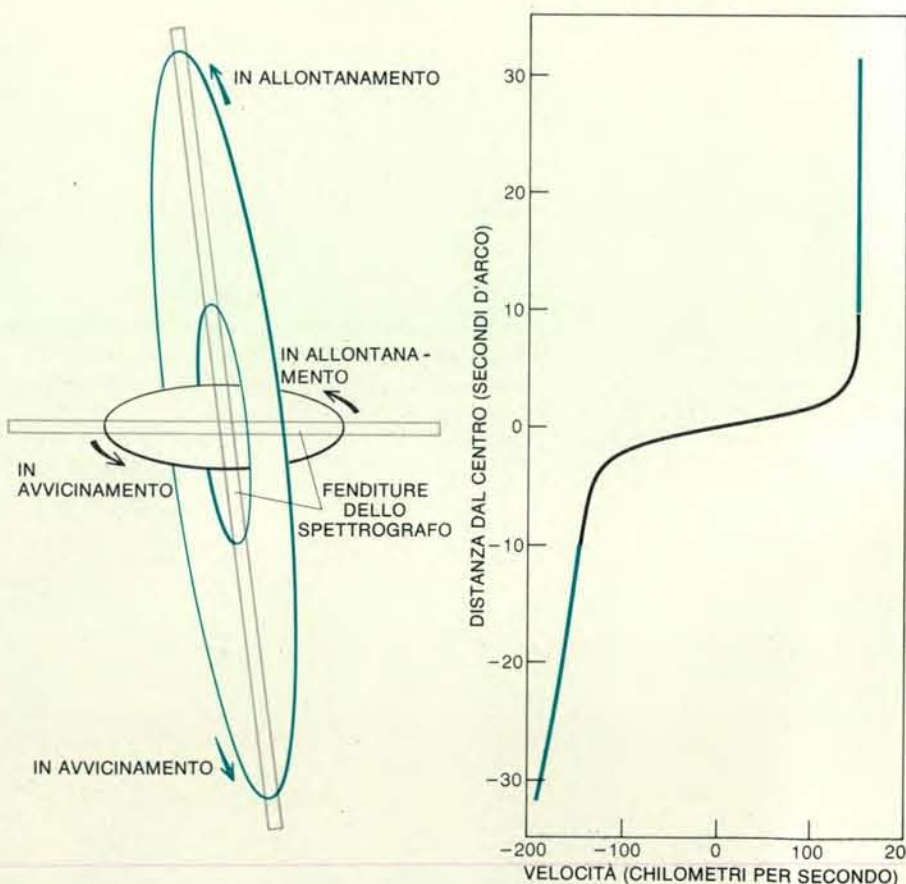
tutte le altre galassie a spirale che io e i miei colleghi abbiamo studiato, trasferisce inoltre il Sole da una posizione piuttosto periferica a una posizione decisamente più centrale.

La prima conclusione di massima che si può ricavare da questi risultati è che, via via che ci si sposta nelle misurazioni dal centro galattico verso l'esterno, la massa complessiva della materia luminosa e di quella oscura presente in ogni punto diminuisce, ma molto lentamente, mentre la luminosità (misurata nella regione blu dello spettro) cala molto rapidamente. Di conseguenza il rapporto tra la densità locale di massa e la luminosità locale (nel blu), che per brevità chiameremo rapporto *M/L*, aumenta regolarmente con la distanza dal centro galattico. Nel nucleo, infatti, una massa relativamente ridotta fornisce una luminosità notevole, mentre a grandi distanze grandi masse producono poca luminosità. Se non esistesse nessuna forma di concentrazione di materia invisibile attorno alle galassie, la distribuzione della massa coinciderebbe semplicemente con quella della luminosità e il rapporto *M/L* resterebbe all'incirca costante su tutto il disco, dal centro fino al bordo.

Misurando massa e luminosità in unità di masse solari e di luminosità solare, il rapporto *M/L* del Sole è ovviamente uguale a 1. In queste unità (trascurando il denominatore che è sempre 1) il valore medio di *M/L* nelle vicinanze del nucleo di una galassia a spirale è vicino a quello del Sole e arriva al massimo a 2 o a 3. Verso il bordo del disco visibile, al calare delle luminosità, *M/L* sale fino a 10 o 20. Oltre il disco visibile, infine, dove la luminosità cade praticamente a zero mentre la massa si mantiene elevata, il valore medio di *M/L* balza fino alle centinaia.

Per tentare di identificare i costituenti dell'alone invisibile è necessario chiedersi quali sono i corpi celesti che hanno valori molto alti di *M/L*. Le stelle simili al Sole sono ovviamente da escludere; le stelle calde, giovani e molto luminose che individuano la struttura dei bracci di una galassia a spirale sono candidati ancora meno soddisfacenti: i loro valori di *M/L* sono intorno a 10^{-4}. Passando all'estremo opposto le nane rosse, molto vecchie, che popolano il rigonfiamento centrale e le regioni esterne di una galassia, hanno una massa piccola e una luminosità nel blu ridotta; i loro valori di *M/L*, pari a circa 20, sono ancora di gran lunga troppo bassi rispetto ai valori necessari per l'alone. Per di più un alone composto di nane rosse di massa particolarmente piccola rivelerebbe la sua presenza con una forte emissione nella regione infrarossa dello spettro. Tutti i tentativi finora compiuti per rivelare la presenza di un alone dal suo irraggiamento nel visibile, nell'infrarosso, o nella regione radio e X sono falliti.

Che altre possibilità rimangono? Le stelle normali irradiano energia prodotta in processi termonucleari che trasformano in elementi più pesanti l'idrogeno e l'elio. Questi processi nucleari possono avere inizio solo in corpi di massa abbastanza grande da far salire la temperatura nel nucleo della stella, per effetto della gravitazione, fino a diversi milioni di kelvin. La massa minima necessaria per questo è pari a circa 0,085 volte la massa del Sole. (Giove, il pianeta più grande del sistema solare, ha una massa che è solo un centesimo di questo valore.) Un alone composto di corpi di tipo planetario, magari protostelle che non si sono trasformate in stelle, è un'ipotesi piuttosto improbabile, ma in linea di principio concepibile. In definitiva l'unica caratteristica nota dell'alone è che deve contenere materia oscura in qualsiasi forma, dai neutrini ai buchi neri, che soddisfi il vincolo imposto dal rapporto *M/L*.

La galassia «anonima» AO 136-0801, così chiamata perché non compare nei normali cataloghi astronomici (i numeri indicano la sua posizione espressa in ascensione retta e declinazione), vanta un vistoso anello attorno al disco. La regione ovale al centro è un disco di stelle in rotazione visto quasi di taglio. Anche le stelle e il gas dell'anello stanno ruotando, ma in un piano quasi perpendicolare a quello del disco. Il disegno al centro indica le direzioni di rotazione e l'orientamento delle fenditure dello spettrografo durante le misurazioni delle velocità orbitali nel disco e nell'anello. Il grafico a destra riporta i risultati delle misurazioni. A dieci secondi d'arco dal centro le velocità nel disco *(in nero)* sono virtualmente uguali a quelle nell'anello *(in colore)*. Nell'anello, però, si riesce a rilevare le velocità fino a quasi tre volte il diametro ottico del disco, e i valori rimangono praticamente costanti. Sembra chiaro che la massa continui ad aumentare linearmente fino a distanze molto superiori al raggio del disco e che i costituenti dell'anello rispondano a un potenziale gravitazionale sferico e non a disco.

Fin qui ho descritto le proprietà rotazionali di galassie a spirale normali e relativamente isolate. Di tanto in tanto, però, la natura si lascia sfuggire un'opportunità eccezionale e imprevedibile di sondare i suoi segreti, e in qualche caso ha offerto ulteriori prove dell'esistenza di valori elevati di *M/L* a grande distanza dal nucleo di altre galassie. Recentemente François Schweizer della Carnegie Institution, Bradley C. Whitmore della Arizona State University e io siamo rimasti profondamente impressionati dalla debole galassia «anonima» AO 136-0801, che fa parte di una classe di galassie fusiformi con anelli polari. Si chiama «anonima» perché non compare nei cataloghi standard di galassie: la sigla indica semplicemente la posizione della galassia nel cielo.

Le nostre osservazioni della distribuzione della luce attraverso il «fuso» indicano che si tratta di un disco di stelle poco luminoso visto quasi di taglio, pressoché privo di gas e di polvere e senza la minima traccia di una struttura a spirale. Queste galassie, che nella classificazione formano il gruppo S0, rappresentano una percentuale abbastanza notevole di tutte le galassie a disco. Con i soliti metodi abbiamo rilevato le proprietà della rotazione del disco misurando lo spostamento Doppler delle righe di assorbimento negli spettri delle stelle che lo compongono. A poca distanza dal centro lungo l'asse maggiore del fuso le velocità di rotazione arrivano a 145 chilometri al secondo, un valore molto vicino a quelli rilevabili in galassie del tipo Sa poco luminose. Lungo l'asse minore, invece, le velocità orbitali non presentano nessuna componente lungo la linea di vista, il che conferma l'ipotesi che si tratti di un disco di stelle in rotazione.

La caratteristica eccezionale di AO 136-0801 è un grande anello, anch'esso visto quasi di taglio attorno all'asse minore del fuso che passa quasi davanti al centro di rotazione del disco *(si veda l'illustrazione nella pagina a fronte)*. L'anello è costituito da gas, polvere, e stelle giovani molto luminose. Il gas rivela la sua presenza con uno spettro di righe di emissione, la polvere grazie agli effetti di assorbimento visibili dove l'anello attraversa frontalmente il fuso, e la componente stellare per il suo aspetto nodoso e bluastro visibile nelle fotografie. Il diametro massimo dell'anello è parecchie volte più esteso dell'asse maggiore del fuso; per questo il moto dei suoi componenti offre un'occasione unica per rilevare il campo gravitazionale di una galassia perpendicolarmente al disco galattico e a distanze maggiori del raggio visibile del disco.

In questo grafico sono riportati i rapporti tra massa e luminosità relativamente ad agglomerati di materia su scale sempre maggiori. Si basa su un altro grafico ideato da Herbert J. Rood dell'Institute for Advanced Study di Princeton. Il valore di *M/L* corrispondente a una densità di materia sufficiente ad arrestare l'espansione dell'universo è pari a circa 700. Per le galassie questo valore è inferiore a 10, ma tende a crescere all'aumentare delle dimensioni dell'agglomerato di materia.

Le nostre osservazioni spettrografiche confermano che l'anello è effettivamente in rotazione in un piano ortogonale a quello del disco. Sembra improbabile che questa configurazione dinamica possa essersi formata nel normale contesto dell'evoluzione di una galassia a disco isolata; deve piuttosto trattarsi del risultato di un evento relativamente raro, come l'incontrò con un'altra galassia o con un disco di gas. In ogni modo misurando lo spostamento delle righe di emissione abbiamo stabilito che la velocità di rotazione dell'anello è di circa 170 chilometri al secondo e che la curva della velocità si mantiene piatta o leggermente in salita fino a una distanza pari almeno al triplo del raggio interno del disco. Tracciando le curve delle velocità nel disco e nell'anello in funzione della distanza in uno stesso grafico si nota che a uguali distanze dal centro della galassia i valori sono quasi identici. L'alta velocità di rotazione dell'anello costituisce un forte indizio della presenza di un alone di massa notevole che si estende almeno tre volte più lontano del raggio visibile del disco. La forma dell'alone, per di più, dev'essere molto più simile a una sfera che a un disco; i calcoli infatti provano che se lo schiacciamento dell'alone fosse pari a quello del disco le velocità sul piano del disco sarebbero dal 20 al 40 per cento maggiori di quelle al di fuori.

Fino a questo punto ho descritto rilevamenti di massa compiuti misurando le velocità di corpi orbitanti di riferimento, esterni alla galassia, di corpi orbitanti nel disco centrale di una galassia e infine di corpi in orbita polare attorno a una galassia insolita. Esistono però anche altri casi particolari che contribuiscono a far luce sulla quantità di materia oscura presente nell'universo. Spesso, infatti, le galassie formano coppie, e in questi casi una delle due si può considerare un corpo di riferimento in orbita attorno all'altra. L'analisi di un sistema di questo tipo è resa particolarmente complicata dal fatto che sia l'orientamento dell'orbita nello spazio, sia la posizione della galassia nell'orbita sono sconosciute. È possibile, però, ricorrere alle proprietà osservate in un grande campione di galassie doppie (la differenza tra le velocità delle due galassie della coppia, la loro distanza angolare e la loro luminosità) per ricavare, attraverso considerazioni di tipo statistico, la distribuzione probabile dei parametri orbitali e dei rapporti *M/L* idonei per le varie galassie.

Diverse analisi condotte indipendentemente da Edwin L. Turner di Princeton, Steven D. Peterson presso la Cornell University, Linda Y. Schweizer della Carnegie Institution e I. D. Karachentsev dell'Osservatorio astrofisico speciale dell'URSS forniscono per *M/L* valori medi compresi tra 25 e 100. Si tratta di valori mediati su una distanza pari alla distanza tra le due componenti di ogni coppia, una distanza uguale di solito a diversi diametri galattici, ossia dell'ordine dei 100 chiloparsec. Questo risultato contribuisce a rafforzare la convinzione secondo cui le galassie possiedono aloni di materia oscura con valori di *M/L* molto elevati e che si estendono molto al di là rispetto ai loro limiti visibili.

Adesso possiamo finalmente tornare alla nostra domanda iniziale: l'universo

contiene abbastanza materia invisibile da far salire la sua densità media a 5×10^{-30} grammi per centimetro cubo, il valore necessario per «chiudere» l'universo e arrestare la sua espansione? Come abbiamo visto questo valore verrebbe raggiunto se la densità della materia oscura arrivasse a superare quella della materia luminosa di circa 70 volte. In alternativa si potrebbe anche esprimere ciò che serve per chiudere l'universo in termini di rapporto tra massa totale e luminosità. Ponendo il valore uguale a 1 per il Sole, il valore medio necessario per chiudere l'universo è pari a circa 700.

C'è qualche motivo per credere che il valore di M/L si avvicini a 700? Facendo una media su tutti i dischi visibili delle galassie a spirale il rapporto tra massa totale (luminosa e non luminosa) e luminosità è circa 5. Nelle galassie di tipo S0 e in quelle ellittiche, però, il risultato è più elevato: dell'ordine di 10. Nelle galassie doppie e nei piccoli gruppi di galassie M/L cresce fino a valori compresi tra 50 e 100. Finalmente, analizzando i moti delle galassie nei grandi ammassi si ottengono per M/L valori nell'ordine delle centinaia. Questa crescita del valore medio di M/L all'aumentare del raggio del sistema è stata messa in evidenza per la prima volta da Einasto, Ants Kaasik ed Enn Saar dell'Accademia delle scienze dell'Estonia e anche da Ostriker e Peebles e da Amos Yahil della State University di New York a Stony Brook. Fino a oggi non vi sono prove dell'esistenza di valori di M/L superiori a quello critico di 700 necessario per chiudere l'universo, ma quelli più elevati tra i valori derivati si avvicinano moltissimo alla soglia. Alcuni fisici, poi, ritengono piuttosto significativo il fatto che i valori stimati sembrino convergere proprio su quello critico invece di mantenersi decisamente più alti o più bassi di qualche ordine di grandezza.

La scoperta del fatto che la distribuzione della luce nell'universo non sia un indice valido della distribuzione della massa rende purtroppo più difficili tutte quelle ricerche che abbracciano distanze enormi e grandi periodi di tempo. Sembra assodato che una percentuale sconosciuta della massa di ogni galassia a spirale, come pure della massa contenuta negli ammassi di galassie, sia nascosta sotto qualche forma non luminosa; è ancora impossibile dire se certe regioni che sembrano non contenere galassie siano vuote solo di luce o anche di massa. Per rispondere a questa domanda gli astronomi dovranno essere abbastanza acuti da escogitare nuove tecniche di osservazione e i fisici dovranno stabilire le proprietà di alcune forme insolite della materia. Solo allora sarà possibile determinare la natura di questa onnipresente materia oscura, stabilire la vera entità della massa e delle dimensioni delle galassie e forse apprendere finalmente quale sia il probabile destino dell'universo.

L'ecologia fisiologica dei cetacei

Questi animali marini, come altri mammiferi, hanno un modo di vita che esige un alto consumo di energia, realizzato sviluppando particolari adattamenti come l'immersione in profondità per lunghi periodi di tempo

di John W. Kanwisher e Sam H. Ridgway

Non ci si rende generalmente conto che 70 specie di mammiferi interamente marine - i cetacei misticeti e odontoceti - adempiono negli ecosistemi marini a una funzione altrettanto essenziale di quella svolta dai mammiferi terrestri sulla terraferma. Il grande successo di questi ultimi alla fine del Mesozoico (circa 65 milioni di anni or sono) va attribuito in gran parte al loro modo di vita a elevato consumo di energia, che comprende aspetti come l'omeotermia e la presenza di un sistema nervoso centrale ben sviluppato. Quando alcune di queste specie terrestri andarono a occupare una nicchia ecologica marina, continuarono a evolvere lo stesso tipo di adattamenti, che comportavano un elevato consumo di energia.

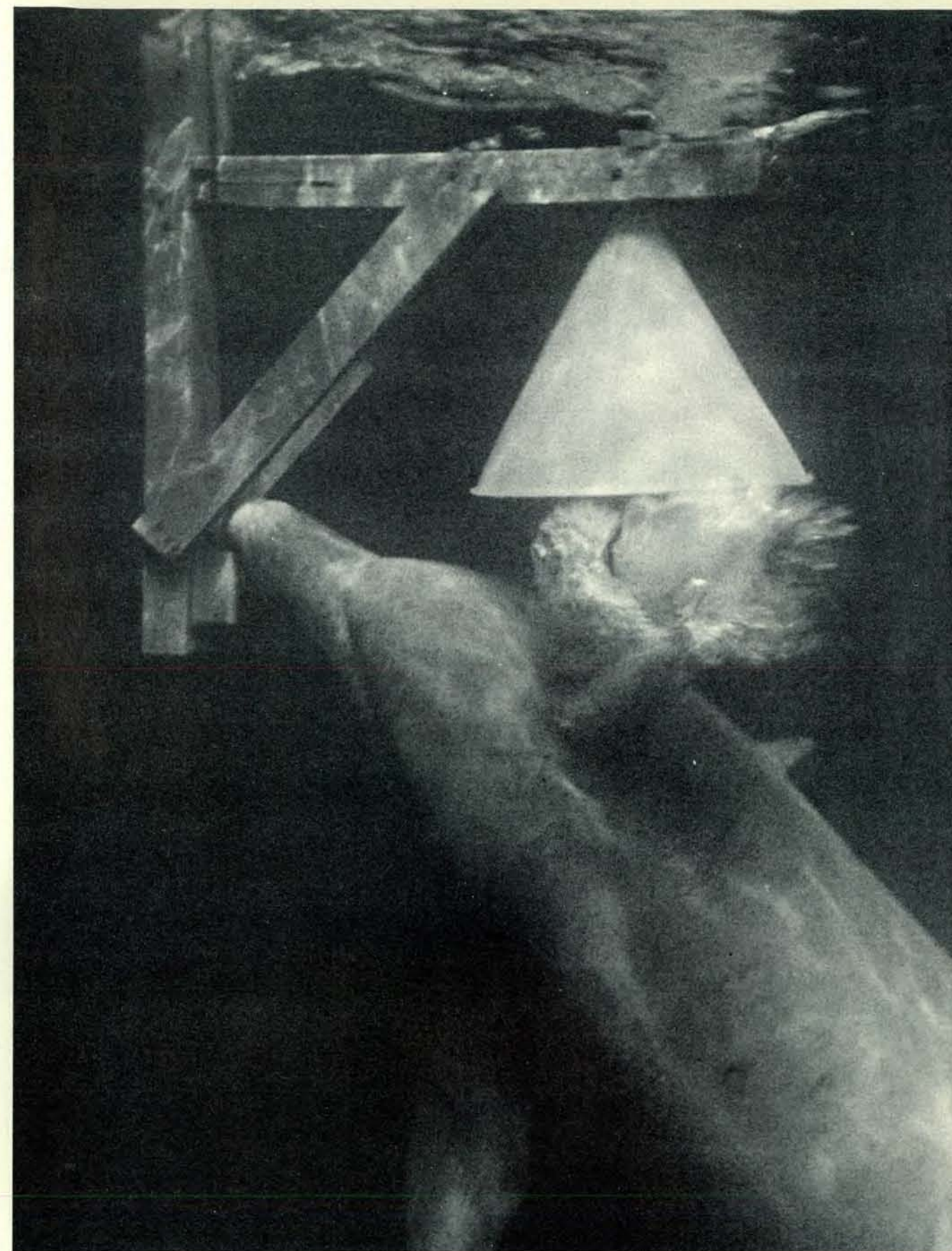
In che modo le balene e le focene - appartenenti all'ordine cetacei - hanno superato i numerosi ostacoli fisiologici posti dalla vita negli oceani? Si pensi, per esempio, al loro ciclo vitale estremamente stressante. Appena espulso dal canale vaginale nell'acqua, a una temperatura talvolta vicina al punto di congelamento, il piccolo deve farsi strada senza aiuto fino alla superficie, ancor prima di inspirare la prima vitale boccata d'aria; per essere allattato dalla madre deve trattenere il respiro e ritornare sott'acqua; quando, infine, è in grado di autoalimentarsi deve sviluppare un repertorio di comportamenti del tutto nuovo, poiché deve cercare le sue prede in immersione. Quando ha raggiunto l'età adulta, deve infine padroneggiare un'altra serie di movimenti ginnastici complessi per potersi riprodurre e far nascere un nuovo individuo, che darà inizio a un nuovo ciclo vitale.

Si consideri anche la sorprendente diversità dei cetacei per quanto riguarda sia la distribuzione geografica sia la mole. I cetacei abbondano in tutti gli oceani, dai tropici ai confini della calotta polare, sia a nord sia a sud. Molte specie migrano al variare delle stagioni dai mari freddi a quelli caldi e viceversa. La più grossa balenottera azzurra adulta è quasi 20 000 volte più pesante del più piccolo nato di focena, eppure nei mari tropicali e in quelli polari si trovano specie sia di grandi sia di piccole dimensioni. I cetacei provvisti di denti - gli odontoceti - comprendono sia piccole focene come la focena comune, a un estremo della loro scala dimensionale, sia il grande capodoglio all'altro estremo. Come mai quest'ultimo è il cetaceo che si immerge a maggiori profondità e può rimanere sott'acqua più a lungo? In che modo riesce a evitare l'embolia gassosa e altri problemi di ordine fisiologico che gli esseri umani incontrano quando si tuffano in acque molto meno profonde? Le focene, essendo più piccole del più piccolo dei loro più grossi cugini odontoceti, presentano un rapporto superficie/volume molto più elevato. Come riescono allora a mantenersi calde nelle acque fredde? Interrogativi di questo tipo hanno motivato la prima di molte nostre indagini: la termofisiologia delle focene.

Le nostre osservazioni su focene in cattività a Marineland in Florida hanno dimostrato che questi animali sono in grado di conservare una temperatura interna stabile, vicina ai livelli umani e altrettanto ben regolata. Questo fatto ci ha messo in imbarazzo, in quanto l'acqua sottrae il calore da un oggetto molto più velocemente dell'aria. Avevamo pensato che questi mammiferi, soggetti a stress termici, dovessero presentare temperature corporee quanto mai variabili, tipiche degli animali ibernanti terrestri. Non essendo questo il caso, abbiamo deciso di quantificare la gravità del problema che deve essere affrontato da una piccola focena che vive nelle acque fredde delle alte latitudini per quanto riguarda la sua termoregolazione.

Dal punto di vista del calcolo del flusso termico, un cetaceo può essere considerato come una massa carnosa uniformemente calda, circondata da uno strato sottocutaneo di grasso isolante. Il flusso di calore verso l'esterno, proveniente dalla massa interna, incontrando la resistenza termica del grasso isolante, deve essere sufficientemente intenso per conservare lo scarto termico fra l'acqua circostante e la massa calda interna. È noto che questa differenza è uguale al prodotto della resistenza termica del grasso per il flusso in uscita del calore metabolico generato dalla massa interna; quando l'isolamento è insufficiente, deve essere prodotta una maggiore quantità di calore corporeo se l'animale deve conservare tale massa a temperatura costante. Per esempio, nel caso di un mammifero terrestre, qualora la temperatura della massa interna si abbassi anche di solo 0,5 gradi centigradi, l'animale comincia a tremare per aumentare il proprio metabolismo. Sui lunghi periodi, squilibri di questo tipo possono essere corretti aumentando il grado di isolamento termico. Il mammifero terrestre produce un mantello invernale e il cetaceo, quando si trova in acque polari, sviluppa uno strato di grasso più spesso.

Quando l'animale conduce vita attiva, un simile incremento della barriera termica può essere di fatto eccessivo. Infatti, quando un cetaceo nuota velocemente, l'aumento dell'attività metabolica fa surriscaldare la sua massa interna; esso deve quindi ricorrere a un espediente di tipo circolatorio: l'aumento del flusso sanguigno in corrispondenza della superficie corporea, particolarmente nelle pinne e negli spiracoli, scavalca dal punto di vista termico lo strato di grasso isolante, riportando la temperatura della massa interna a livelli normali. I cetacei raggiungono generalmente una temperatura corporea costante principalmente in virtù di queste variazioni del flusso sanguigno. Questo vale anche per gli esseri umani: quando hanno troppo caldo, viene dirottata una maggior quantità di sangue verso la superficie del corpo, producendo l'aspetto avvampato caratteristico della persona accaldata, la cui epidermide diventa invece violacea se esposta al freddo intenso.

Vi sono, però, differenze fondamentali fra la fisiologia degli esseri umani e quella dei cetacei. Per esempio, i cetacei non hanno ghiandole sudoripare: il raffreddamento per evaporazione è impossibile nell'ambiente acquatico; parallelamente, per

Questo tursiope troncato, addestrato alla esecuzione di compiti in ambiente subacqueo, ha collaborato con gli autori negli esperimenti sulla fisiologia in apnea. Noto con il nome di Tuffy, questo piccolo delfino si immergeva sollecitato da un segnale sonoro e, a intervalli diversi, rispondeva a un secondo segnale, premendo un pulsante e espirando in un imbuto, in modo che si potesse misurare la sua riserva residua di ossigeno.

quanto riguarda l'isolamento superficiale, gli esseri umani hanno uno strato adiposo sottocutaneo veramente insignificante. I nuotatori subacquei sono riusciti a imitare i cetacei indossando tute di gomma espansa ai fini dell'isolamento; sulla terraferma, gli uomini hanno eliminato lo svantaggio della loro origine essenzialmente tropicale coprendosi, fra le altre cose, con il materiale isolante invernale prodotto da altri mammiferi terrestri.

Ancora prima di dare inizio alla nostra indagine, eravamo perfettamente consci che i cetacei di minori dimensioni sono quelli che incontrano i più seri problemi per conservare la normale temperatura corporea dei mammifero. Considerando l'area superficiale totale che cede costantemente calore all'acqua circostante, quanto più l'animale è piccolo, tanto minore è il volume della sua massa interna che genera il calore metabolico necessario per mantenere lo scarto di temperatura all'interno della barriera formata dallo strato di grasso. Oltre alla sfavorevole geometria dei rapporti superficie/volume, anche le misure in scala lineare depongono negativamente nei confronti di qualsiasi piccola focena che cerchi di conservare il proprio calore in acque fredde. A causa delle loro ridotte dimensioni, gli animali più piccoli tendono a presentare uno strato di grasso più sottile.

Indipendentemente dalla loro mole, i cetacei mantengono la normale temperatura interna dei mammiferi, che è di circa 37 gradi centigradi. Nell'Artide, dove la temperatura dell'acqua può raggiungere i −2 gradi centigradi, la differenza di temperatura attraverso la barriera sottocutanea è di circa 40 gradi. Ai Tropici, dove la temperatura dell'acqua può essere attorno ai 30 gradi centigradi, il gradiente termico attraverso la barriera rappresenta solo un quinto di quello riscontrato nell'Artide. Pertanto, una focena adattata alla vita in acque calde sembrerebbe destinata a incontrare serie difficoltà di ordine termico quando si trasferisce nei mari polari. Eppure la focena comune *(Phocaena phocaena)*, uno dei cetacei di minor mole, si trova soprattutto nelle acque fredde boreali.

Per determinare le dimensioni del problema affrontato dalle focene comuni in acque fredde, abbiamo calcolato la quantità di calore metabolico che dovrebbero generare per mantenere la temperatura giusta. Per determinare l'area superficiale totale e lo spessore del loro strato di grasso isolante, abbiamo effettuato misurazioni dirette su un esemplare di focena comune, arenatosi su una spiaggia. Dopo aver introdotto i nostri dati nella formula relativa al calcolo della perdita di calore, l'unica conclusione che ne abbiamo tratto è stata che una focena così piccola, con il normale apporto di calore metabolico dei mammiferi, non poteva mantenersi calda nelle acque nordiche. Eppure, con i nostri occhi, vedevamo le focene che sguazzavano nelle stesse acque gelide che avrebbero dovuto ucciderle.

Un paradosso di questo genere disturba sempre i biologi, abituati a prevedere regolarità quando controllano un parametro comune a gruppi animali diversi. La quantità di calore metabolico prodotto da animali di dimensioni diverse è un dato di confronto che, fino a quel momento, sembrava abbastanza consolidato. Per esempio, da tempo i tassi metabolici di numerosi mammiferi terrestri venivano ordinati nella cosiddetta curva «dal topo all'elefante». Si era sempre creduto che, dato il peso corporeo dell'animale, sulla base della curva che lo rappresenta, fosse possibile calcolare con buona approssimazione il calore metabolico prodotto da un mammifero qualsiasi. La nostra valutazione del tasso metabolico necessario a mantenere calda una focena in acqua fredda era però così elevata da contrastare chiaramente con i dati relativi ai mammiferi terrestri.

Vi erano solo due modi per spiegare questo paradosso. Il primo, molto semplicemente, presupponeva che il metabolismo dei cetacei non si conformasse alla curva «dal topo all'elefante», costruita per i mammiferi terrestri. Il che voleva dire che i cetacei potevano presentare un metabolismo basale intrinsecamente più elevato che permetteva anche a un esemplare di piccole dimensioni di sopravvivere in acque fredde. La seconda ipotesi era che i cetacei avessero un tasso basale di produzione calorica «normale» (ossia terrestre), ma che generassero, per il fatto di mantenersi costantemente in movimento quando frequentavano le acque fredde, una sufficiente quantità di calore aggiuntivo per combattere il freddo eccessivo. L'unica possibilità di scegliere fra queste due spiegazioni consisteva nell'«interrogare» direttamente le focene. Quale strategia termica aveva stabilito per loro l'evoluzione?

Con la preziosa collaborazione di Gunnar Sundnes, un fisiologo del Havforskningsinstitutt di Bergen, abbiamo organizzato un esperimento sulla perdita di calore. Per prima cosa, abbiamo fatto in modo di mantenere alcune giovani focene in una grande vasca, in condizioni di cattività più o meno sopportabili, trasferendole poi, ai fini sperimentali, in una vasca d'acqua marina più piccola, che è servita da calorimetro per la totalità del corpo. Il piccolo della focena comune pesa solo 15 chili; nonostante che all'inizio il suo strato di grasso sottocutaneo abbia uno spessore di un solo centimetro, la differenza fra temperatura corporea normale e temperatura dell'acqua marina circostante può raggiungere anche i 25 gradi centigradi.

La focena sottoposta all'esperimento rimaneva tranquilla nel contenitore, mentre il calore metabolico che produceva passava attraverso lo strato sottocutaneo, riscaldando l'acqua circostante. Siamo poi riusciti a introdurre nella vasca un termistore e, con una leggera amplificazione, abbiamo potuto produrre una curva, che riportava la temperatura in funzione del tempo, su un registratore su carta con sensibilità fino a 0,001 gradi centigradi. Tenuto conto del volume d'acqua esistente nel contenitore, l'andamento della curva ha fornito una registrazione costante della quantità di calore metabolico che la focena perdeva in un momento qualsiasi.

I risultati sono parsi chiaramente orientati in favore della prima delle due alternative: le piccole focene mostravano un tasso metabolico due o tre volte maggiore rispetto a quello dei mammiferi terrestri dello stesso peso. Questo incremento di produzione termica non è gratuito: alla focena va somministrata una quantità di cibo tre volte maggiore di quella necessaria a un mammifero terrestre dello stesso peso e deve avere una frequenza respiratoria tripla rispetto a quella di animali con tasso metabolico terrestre.

Sulla base dei dati ricavati per le focene, abbiamo continuato a elaborare calcoli termici per cetacei di maggiori dimensioni allo scopo di determinare con quanta maggiore facilità un grosso mammifero, dotato di una geometria più favorevole, riesca a mantenere elevata la propria temperatura corporea in acqua fredda. Abbiamo dapprima considerato una balena teorica, con peso mille volte superiore a quello di una focena comune; dato che i due mammiferi marini hanno una forma simile, le dimensioni lineari della balena devono essere decuplicate. La balena teorica dispone, per generare calore metabolico, di una quantità di tessuto 1000 volte superiore, ma il calore si trasferisce all'acqua circostante attraverso una superficie che è solo 100 volte maggiore di quella della focena.

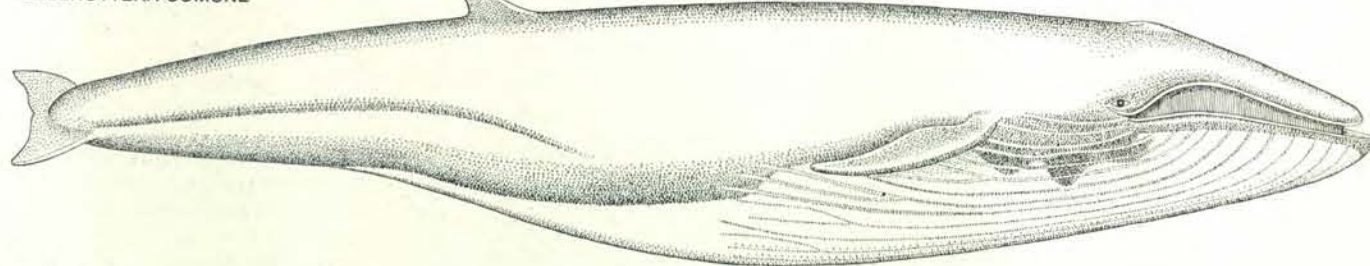
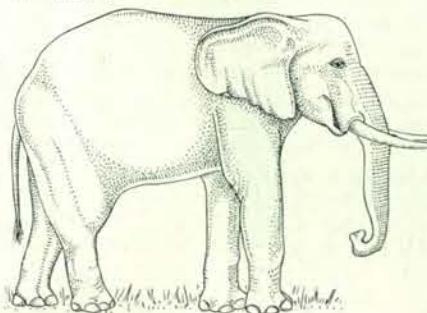
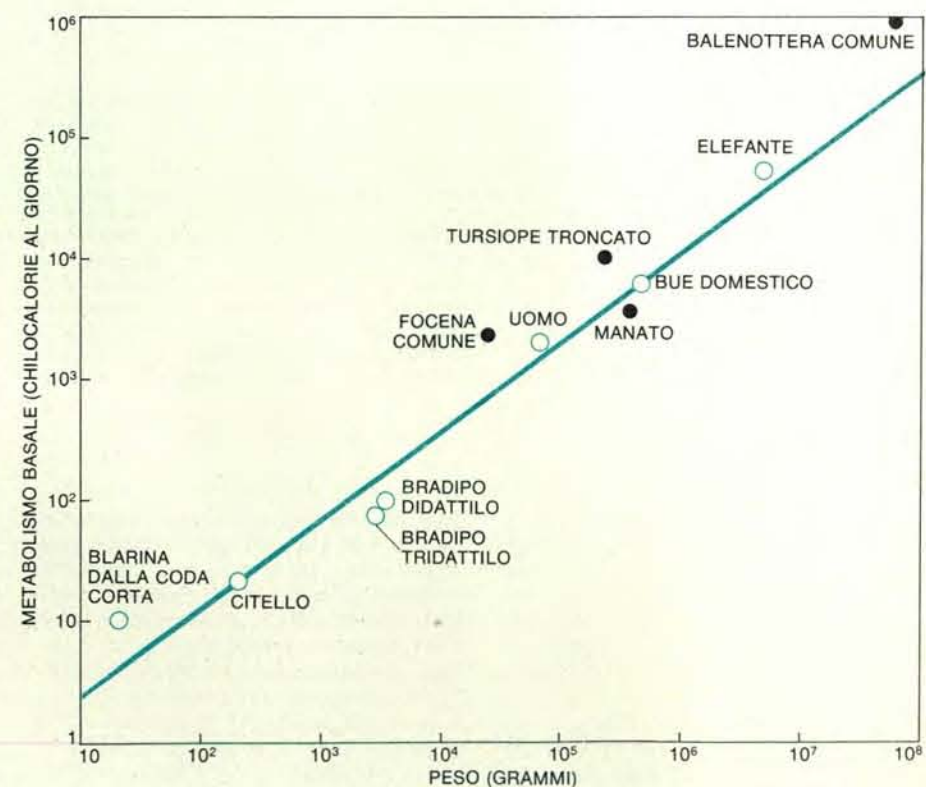
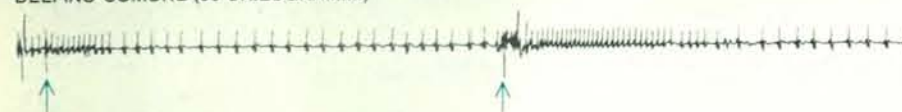
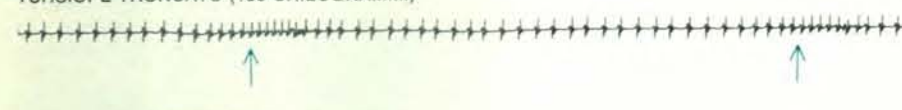
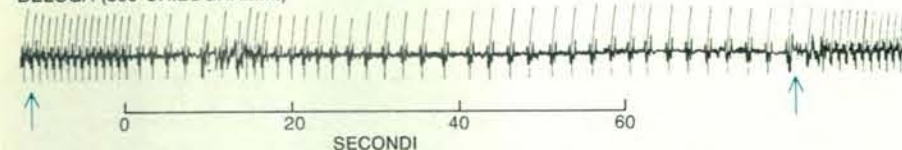
Si pensi a questa minor perdita di calore unicamente in termini di isolamento da parte dello strato di grasso sottocutaneo. Mentre la piccola focena può riequilibrare la propria perdita di calore mediante uno strato dello spessore di due centimetri, la balena, che pesa mille volte di più, allo stesso tasso metabolico della focena dovrebbe ottenere lo stesso risultato con uno strato sottocutaneo di spessore non superiore ai due millimetri. In realtà, invece, una balena di queste dimensioni avrebbe uno strato di grasso di quasi 20 centimetri e, persino in acque polari, è in grado di affrontare problemi di raffreddamento. Un cetaceo così grande risulterebbe 100 volte superisolato.

Se lo strato di grasso sottocutaneo in un grosso cetaceo è più spesso di quanto richiedano le esigenze di isolamento termico, a quali altre funzioni esso adempie? Per prima cosa, il tessuto adiposo, che ha un basso peso specifico, dà un sufficiente galleggiamento che controbilancia la spinta idrostatica negativa dovuta ai muscoli e allo scheletro della balena. In secondo luogo, lo strato di grasso costituisce una riserva nutritiva abbastanza consistente per provvedere al fabbisogno metabolico della balena nel corso delle sue migrazioni stagionali, quando deve sovente attraversare migliaia di chilometri di acque con scarso cibo.

In base alla regola generale che rapporta il tasso metabolico dei mammiferi alla loro mole, si dovrebbe presumere che una balena presenti una produzione calorica per chilogrammo di tessuto considerevolmente più bassa di quella rilevata sperimentalmente nelle focene comuni. Non riuscivamo a pensare a un modo di verificare quest'ipotesi misurando direttamente il tasso metabolico di un animale di 40 tonnellate; pertanto, dovevamo limitarci a valutare la produzione calorica della balena unicamente in base a considerazioni d'ordine anatomico.

A Steinshamn, una stazione baleniera sulle coste della Norvegia, abbiamo proceduto alla dissezione dei polmoni di una balenottera comune *(Balaenoptera physalus)* lunga 14 metri circa. Il volume polmonare della balenottera è stato determinato insufflando dapprima aria compressa nei polmoni e poi svuotandoli attraverso uno spirometro. Abbiamo supposto che il volume di aria espulsa fosse pari al volume d'aria introdotto nei polmoni con ogni singola inspirazione. Attraverso osservazioni in mare, abbiamo determinato la frequenza delle inspirazioni di una balenottera comune. L'immersione normale di questa specie comporta un ritorno alla superficie per cinque rapide inspirazioni e una permanenza in profondità della durata di cinque minuti; una balenottera comune adulta respira, così, a intervalli di un minuto circa in media.

Combinando il volume polmonare e la frequenza di respirazione, abbiamo potuto valutare la quantità d'aria inspirata da una balenottera comune; essa è di 2000 litri al minuto. Ora, i mammiferi utilizzano in generale circa il 6 per cento dell'ossigeno atmosferico che respirano (il 21 per cento della miscela atmosferica totale di azoto, ossigeno e altri gas). Dato che ogni centimetro cubo di ossigeno consumato dall'animale produce cinque calorie, siamo stati in grado di fare una stima, anche se approssimativa, del tasso metabolico della balenottera comune; come nel caso della focena comune, il risultato ottenuto ha superato in misura significativa quello della curva relativa ai mammiferi terrestri. Sulla base del semplice raffronto del peso, il tasso metabolico della balenottera comune è persino inferiore a quello dei bradipi, che presentano un tasso conforme ai loro movimenti letargici.

Quando si considerino sia il minor tasso metabolico assoluto dei grossi cetacei sia la loro capacità di accumulare riserve nutritive sotto forma di grasso sottocutaneo, è evidente che tutti questi animali sono relativamente indipendenti dalle fonti alimentari locali e possono soddisfare le proprie esigenze energetiche per tre o sei mesi, unicamente sfruttando la metà della loro provvista di grasso. Questo permette loro di nutrirsi nei ricchi mari polari e di partorire i

La curva «dal topo all'elefante», rappresentata da una linea continua *(in colore)* nel grafico a destra, è stata elaborata negli anni trenta da Francis G. Benedict della Carnegie Institution di Washington come metodo approssimativo per valutare il metabolismo basale di vari mammiferi, espresso in chilocalorie al giorno, sulla base del 75 per cento del loro peso. I cerchi in colore mostrano come alcuni mammiferi terrestri selezionati si conformino bene a questa curva, costruita su un'ordinata e un'ascissa in scala logaritmica. (L'ordinata ha un valore massimo di un milione di chilocalorie e l'ascissa arriva a 100 milioni di grammi). Tutte le misurazioni dirette del metabolismo delle focene comuni *(Phocaena)*, dei tursiopi *(Tursiops)* e la stima basata sul volume polmonare *(pallini in nero)*, della balenottera comune *(Balaenoptera)*, si collocano ben al di sopra dei valori per i mammiferi terrestri, anche se a parità di peso il loro metabolismo è inferiore a quello dei mammiferi terrestri pigri quali i bradipi.

I battiti cardiaci degli odontoceti sono stati registrati sia mediante telemetria in tursiopi come quello che appare nella fotografia in alto, al quale è stato applicato un trasmettitore ultrasonico, sia per misurazione diretta tramite elettrocardiogrammi come per gli altri due tracciati, che mostrano il battito cardiaco di un delfino comune *(Delphinus delphis)* di 50 chilogrammi e di un beluga o delfinattero bianco *(Delphinapterus leucas)* di 500 chilogrammi. Le frecce in colore indicano l'inizio dell'inspirazione caso per caso. Il battito cardiaco dei tre animali diventava più frequente ad ogni inspirazione, atto mirante ad accelerare l'immissione di ossigeno. In seguito, esso decelerava fino a raggiungere il livello basale, sia che gli animali si immergessero sia che riposassero in superficie o fossero addirittura fuori dall'acqua. Esperimenti precedenti (condotti dal fisiologo P. F. Scholander su foche e anatre in cattività) avevano dimostrato una radicale diminuzione della frequenza del battito cardiaco a seguito di immersione forzata. I cetacei sopra menzionati, invece, non hanno fatto registrare questa diminuzione. I risultati ottenuti fanno pensare che la reazione degli animali sotto costrizione fosse di allarme, dato che venivano immersi contro la loro volontà.

loro piccoli a migliaia di chilometri di distanza nelle calde acque tropicali.

In confronto, la piccola focena comune soffre di restrizioni molto maggiori. Il suo tasso metabolico più alto rende più pressanti le esigenze nutritive e limita il tempo a disposizione per le migrazioni; inoltre, rende più difficoltoso trattenere il respiro: quando la focena s'immerge in cerca di cibo, deve dipendere dalla quantità di ossigeno che può inspirare. Le nostre successive indagini si sono concentrate appunto su come questo animale riesce a operare con l'ossigeno di cui può disporre.

Sapevamo già che il modo di vita a elevato dispendio energetico dei mammiferi è stato reso possibile da un altrettanto elevato consumo d'ossigeno; a sua volta, questo fenomeno era dipeso dallo sviluppo evolutivo parallelo di sistemi respiratori e cardiovascolari più efficienti di quelli degli anfibi e dei rettili. Quando i mammiferi terrestri antenati dei cetacei abbandonarono la terraferma (e la sua illimitata disponibilità di ossigeno) preferendo la vita acquatica, si rese indispensabile un ulteriore sviluppo evolutivo che favorisse il trasporto e l'immagazzinamento di maggiori quantità di ossigeno. In che modo questo obiettivo venne raggiunto?

Consideriamo il capodoglio, il cetaceo che presenta le più impressionanti capacità d'immersione. Mentre la focena può rimanere sott'acqua solo per pochi minuti, le immersioni del capodoglio possono protrarsi anche per un'ora, eppure anche la modesta prestazione subaquea della focena ridicolizza l'uomo che si immerge in apnea. Effettivamente, sia i grossi sia i piccoli cetacei possono trasportare in profondità molto più ossigeno di un uomo.

Uno dei fattori di questa prestazione consiste nel volume sanguigno considerevolmente maggiore nel caso dei cetacei: una quantità di sangue che, per unità di peso corporeo, supera di due o tre volte quella degli esseri umani, con una capacità di trasporto dell'ossigeno alquanto superiore. I cetacei dispongono, inoltre, di un ulteriore quantitativo di ossigeno legato alla mioglobina presente nei loro muscoli. La mioglobina (una proteina che è, in realtà, una subunità dell'emoglobina dei globuli rossi del sangue, la proteina che veicola l'ossigeno) è responsabile del caratteristico colore rosso scuro dei muscoli. Nei muscoli di alcune specie di cetacei, la quantità di mioglobina in effetti è sufficiente a trasportare una quantità di ossigeno supplementare, corrispondente a oltre la metà di quella contenuta nei globuli rossi del sangue.

Quando un cetaceo emerge dopo una prolungata immersione e ha esaurito ambedue queste riserve di ossigeno, l'elevata efficienza del suo sistema vascolare fa sì che il sangue passi rapidamente attraverso i polmoni per rifornirsi di ossigeno; a ciò contribuiscono un cuore particolarmente possente e un sistema circolatorio caratterizzato da vasti reticoli di capillari (le cosiddette «reti mirabili») e da grandi seni venosi. L'esatta funzione di queste insolite strutture può essere solo intuita, ma è probabile che esse facilitino un rapido ricarico d'ossigeno nei globuli rossi. La rapidità della circolazione sanguigna si riflette nel fatto che, mentre l'animale sta respirando in superficie, il suo battito cardiaco è più frequente. Questa è una differenza fondamentale nella respirazione fra cetacei e mammiferi terrestri, la cui respirazione è abitualmente regolare e il flusso sanguigno relativamente stabile.

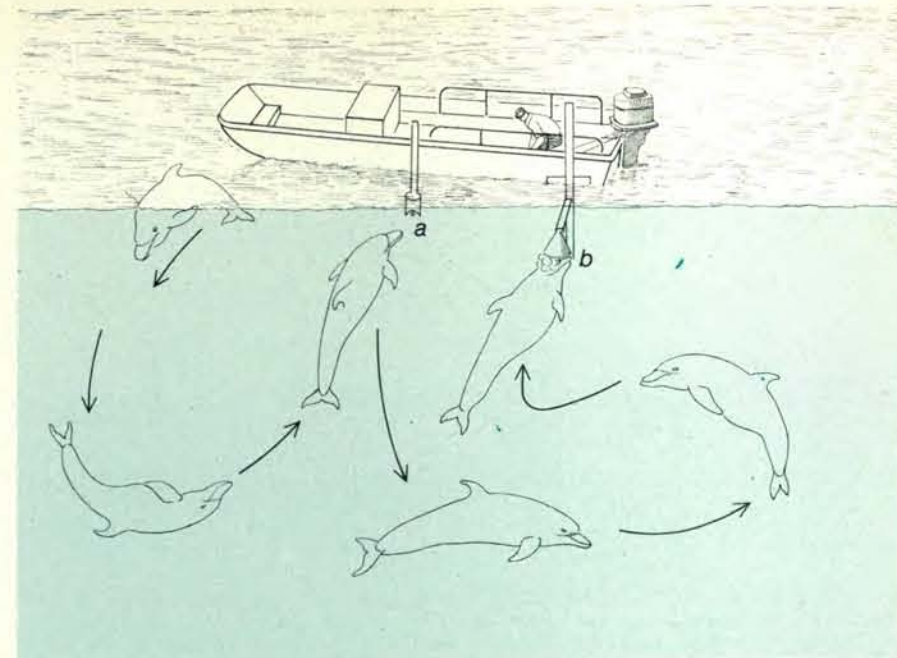
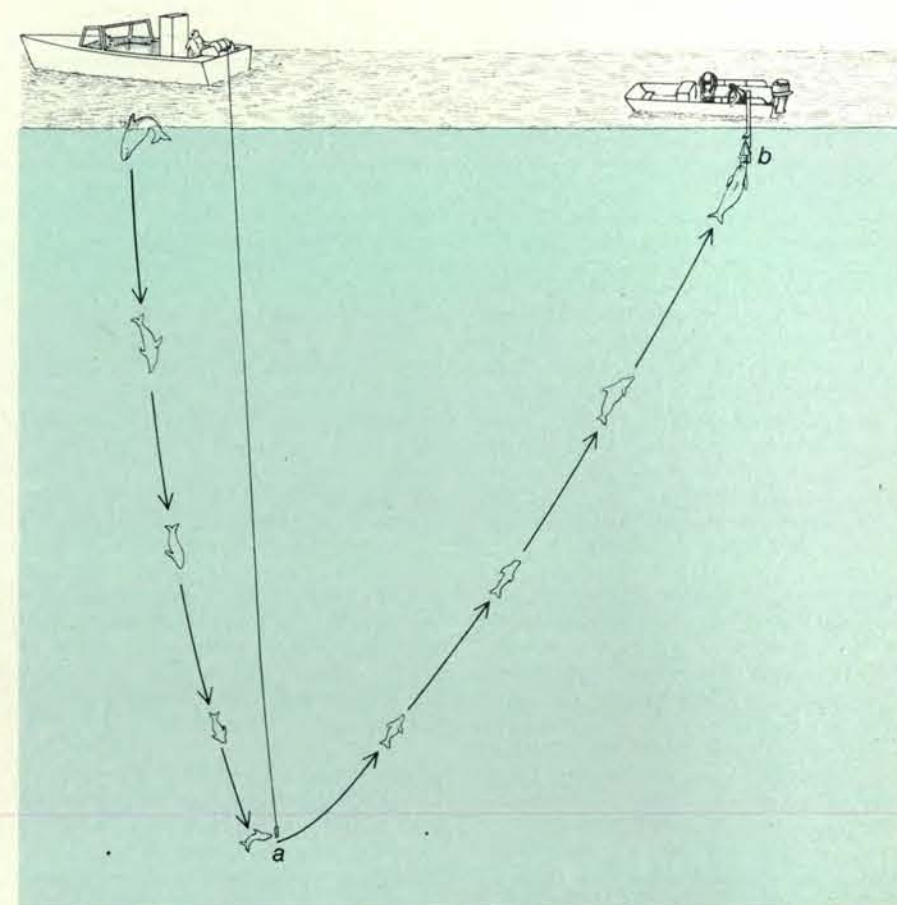
Nel frattempo i globuli rossi ricaricati trasportano parte della loro provvista di ossigeno alla mioglobina delle cellule muscolari, che ne è rimasta priva; il trasferimento è agevole poiché la mioglobina presenta una tensione di carico dell'ossigeno molto minore dell'emoglobina; essa lega infatti l'ossigeno più strettamente dell'emoglobina, impedendogli di rientrare nel sistema circolatorio. Conviene considerare l'ossigeno legato alla mioglobina come una risorsa puramente locale, sfruttabile dal muscolo nel corso di un'immersione; per la maggior parte, esso si trova a una distanza dai mitocondri delle cellule muscolari, gli organelli intracellulari al cui interno si svolge la respirazione, che è facilmente superabile per diffusione.

La mioglobina è presente anche nei muscoli dei mammiferi terrestri, ma in quantità molto minori; una delle sue funzioni, in questi animali, è quella di fonte d'ossigeno locale nel corso della contrazione muscolare, quando la pressione intramuscolare tende a frenare il flusso sanguigno nei capillari, interrompendo il normale apporto di ossigeno da parte dei globuli rossi. È evidente che la maggiore quantità di ossigeno trasportata dalla mioglobina dei cetacei è un altro esempio di adattamento all'immersione, sviluppato per accentuazione di una capacità preesistente nei mammiferi.

Quando abbiamo iniziato il nostro studio sulle facoltà subacquee dei cetacei, esisteva già una spiegazione ampiamente accettata dei lunghi tempi d'immersione, caratteristici degli animali acquatici. Essa derivava dagli esperimenti condotti da P. F. Scholander della Scripps Institution of Oceanography, i quali dimostravano che un animale in stato di cattività (un'anatra o una foca) quando viene forzosamente immerso mostra un'improvvisa diminuzione del battito cardiaco per cui il flusso sanguigno così ridotto viene incanalato in modo tale da interrompere in gran parte l'irrorazione sanguigna nei muscoli e nei visceri. Questa ridotta gettata cardiaca porta quasi interamente il sangue al cervello e al cuore, organi che si ritiene abbiano la necessità di un rifornimento costante di ossigeno. (Per esempio, l'interruzione per qualche secondo del flusso sanguigno verso il cervello provoca nell'uomo una perdita di coscienza.) I muscoli depauperati dell'ossigeno, negli esperimenti di Scholander, potevano così contare unicamente su una via energetica anaerobica, con conseguente accumulo di acido lattico quale prodotto collaterale.

Nei normali libri di testo, questa sequenza logica di eventi è citata per spiegare la capacità d'immergersi per lunghi periodi di tempo non solo delle anatre o delle foche, ma anche di altri uccelli acquatici e dei mammiferi marini in generale. Non esistevano, però, praticamente dati fisiologici riguardanti i cetacei. In realtà, l'unica volta che una focena in stato di cattività era stata privata dell'aria in questo modo, a titolo sperimentale, subì un collasso e morì. Per verificare quest'ipotetica reazione all'immersione, decidemmo quindi di raccogliere dati basandoci su cetacei in libertà.

Come primo passo abbiamo applicato dei trasmettitori ultrasonici alle focene, in modo da poterne registrare il battito cardiaco durante il nuoto. Siamo stati sorpresi, innanzitutto, di non trovare conferma all'improvvisa caduta del battito cardiaco al momento dell'immersione, caratteristica delle anatre e foche fatte immergere forzatamente. Per tutto il periodo di immersione, le focene sembravano conservare un livello di circolazione generale relativamente alto; pertanto ci siamo messi d'impegno per misurare più direttamente l'intervallo di tempo loro necessario per esaurire la provvista di ossigeno mentre si trovavano sott'acqua.

Per nostra fortuna, in quel periodo la Marina statunitense stava allenando dei tursiopi troncati *(Tursiops truncatus)* per utilizzarli quali portaordini per i subacquei di grandi profondità. Un istruttore della Marina, William Scronce, s'incaricò di insegnare a uno di questi delfini, di nome Tuffy, a espirare su suo comando all'interno di un imbuto sommerso, in modo

Il delfino Tuffy, un tursiope troncato addestrato dalla Marina statunitense, aveva imparato a immergersi come si nota nello schema, al segnale impartito da un cicalino *(a)* e a rimanere sott'acqua in attesa di un secondo segnale *(b)*. A questo punto, nuotava verso il secondo cicalino, lo toccava col muso ed espirava in un imbuto. Alla fine dell'esercizio veniva ricompensato con tre pesci. L'esperimento ha dimostrato che Tuffy era in grado di rimanere comodamente in apnea per sei minuti, anche se entro i primi tre minuti aveva esaurito quasi tutta la sua provvista di ossigeno.

La capacità di immersione in profondità del tursiope è stata saggiata in un secondo esperimento. Dapprima sono stati calati a 150 metri di profondità un cicalino e un pulsante *(a)*. Al suono del cicalino, Tuffy si tuffava e premeva l'interruttore; poi ritornava senza difficoltà all'imbarcazione per azionare un secondo cicalino *(b)* e ricevere il suo compenso. Quando doveva premere il pulsante alla profondità di 300 metri era costretto a rimanere più a lungo in superficie per recuperare.

Il terzo esperimento mirava a stabilire se un tursiope può essere soggetto a embolie (le gravi conseguenze dell'assorbimento dell'azoto atmosferico da parte del sangue in una immersione prolungata). Al segnale, il tursiope si tuffava ripetutamente a una profondità di 100 metri, rimanendo sott'acqua alla profondità media di 50 metri per 20 minuti ogni ora. Nuotava poi verso un natante collocato presso un'imbarcazione su cui era stato installato uno spettrometro di massa e i ricercatori gli introducevano nella muscolatura, attraverso lo strato sottocutaneo, un ago per prelievi di campioni gassosi. La muscolatura risultò satura d'azoto, a livelli pericolosi per un uomo. Gli autori attribuiscono all'adattamento all'immersione la resistenza del tursiope alle embolie.

che potessimo raccogliere l'aria espirata per analizzarla. Il segnale d'azione per Tuffy era costituito da un cicalino posto sott'acqua: se Tuffy spingeva il pulsante colla punta del muso ed espirava nell'imbuto, riceveva in premio tre pesci. Un altro cicalino dava a Tuffy il segnale per tuffarsi. Variando l'intervallo fra i suoni del cicalino «per l'immersione» e del cicalino «per l'espirazione», fummo in grado di ottenere campioni d'aria distribuiti lungo un ampio arco di tempi d'apnea. Tuffy poteva facilmente trattenere il respiro anche per sei minuti, ma dopo un tempo molto più breve cominciava a dimostrarsi sempre più impaziente mentre era in attesa di udire il suono del cicalino «per l'espirazione»; volgendo lo sguardo in alto verso di noi, digrignava persino i denti.

Studiando i dati forniti da Tuffy siamo stati sorpresi nel constatare con quanta rapidità calava la concentrazione d'ossigeno nei suoi polmoni; per esempio, dopo soli tre minuti, il gas all'interno dei polmoni di Tuffy, dal livello normale atmosferico del 21 per cento, era caduto a meno del 2 per cento; questo significava che, nel corso dei successivi tre minuti di queste «comode» immersioni, l'animale era praticamente in anaerobiosi. Dopo aver trattenuto il respiro per un periodo di tempo così lungo, Tuffy non ripeteva l'operazione prima di aver passato quattro o cinque minuti in superficie respirando velocemente; ed è chiaro che lo faceva per recuperare il debito d'ossigeno. Invece, dopo immersioni di due minuti o meno, si rituffava volentieri anche dopo un breve intervallo passato respirando in superficie.

A questo punto eravamo ormai in grado di capire perché non eravamo riusciti a rilevare un'improvvisa diminuzione dei battiti cardiaci quando le nostre focene, munite di trasmettitori ultrasonici, si tuffavano; al pari di Tuffy, esse avevano compiuto solamente uno sforzo parziale, o nessuno sforzo del tutto, per conservare la riserva iniziale d'ossigeno: quella riserva veniva rapidamente consumata e si produceva immediatamente, in tutti i loro tessuti, cervello compreso, un debito d'ossigeno. Osservammo anche che tutti gli animali acquatici che respirano aria (uccelli tuffatori, foche e manati, oltre ai cetacei da noi presi in considerazione) cercavano di evitare questo debito, limitando le loro immersioni al tempo strettamente necessario, per non esaurire la loro provvista di ossigeno. Per esempio, quando Tuffy e i suoi compagni delfini nuotavano senza essere sollecitati da comandi, le immersioni per periodi di tempo superiori ai due minuti e mezzo erano l'eccezione più che la regola. Se erano liberi di immergersi per intervalli di tempo più brevi, allora potevano continuare a tuffarsi per ore.

Quando un nuotatore subacqueo rimane sott'acqua per un periodo di tempo sufficientemente lungo perché una notevole quantità di azoto gli entri in soluzione nel sangue e risale poi in superficie troppo rapidamente, l'azoto forma bollicine che provocano conseguenze gravi (emboli) o addirittura mortali. Come fanno, allora, le nostre focene, per non parlare dei grossi misticeti che possono immergersi a profondità superiori ai 1000 metri per più di un'ora, a evitare le embolie? Allo scopo di risolvere questo e altri interrogativi correlati, abbiamo predisposto alcuni esperimenti d'immersione a grande profondità per il nostro tursiope addestrato.

Facevamo scendere un cicalino, controllato dalla superficie, verso uno qualsiasi dei punti predisposti a varie profondità e Tuffy era stato istruito a tuffarsi per raggiungerlo quando suonava e ad agire sull'interruttore premendo col muso sul pulsante; quando spegneva il cicalino «per l'immersione», Scronge attivava immediatamente quello «per l'espirazione», che si trovava in superficie. Tuffy risaliva velocemente verso questo secondo cicalino ed espirava all'interno dell'imbuto di raccolta.

Secondo la letteratura scientifica, gli odontoceti non si immergono a grandi profondità. Quando abbiamo iniziato i nostri esperimenti, abbiamo fatto scendere il cicalino «per l'immersione» fino a 150 metri di profondità; Tuffy riusciva comodamente a raggiungere questo limite, riemergeva per fornire il suo campione di gas e, talvolta, esigeva il suo compenso di tre pesci, che gli spettava, ancora prima di inspirare nuova aria: chiaramente non aveva ancora raggiunto il limite estremo di profondità. Quando abbiamo abbassato il cicalino fino a 300 metri, è risultato evidente che Tuffy recuperava molto più lentamente; era chiaro che si trovava molto vicino al limite critico. (Gli istruttori della Marina, alle Hawaii, hanno poi allenato una specie affine a scendere fino a quasi 600 metri.)

Sembrava che Tuffy non risentisse dell'aumento della pressione idrostatica (corrispondente a 30 atmosfere), che si aveva alle maggiori profondità. In presenza di una simile pressione esterna c'era da aspettarsi che i polmoni gli collabissero quasi completamente. Applicando un apparecchio fotografico munito di flash al cicalino in profondità, siamo riusciti a registrare l'aspetto di Tuffy mentre premeva l'interruttore a 300 metri e abbiamo potuto rilevare che effettivamente, la sua cassa toracica era molto deformata dalla pressione idrostatica che agiva sui suoi polmoni.

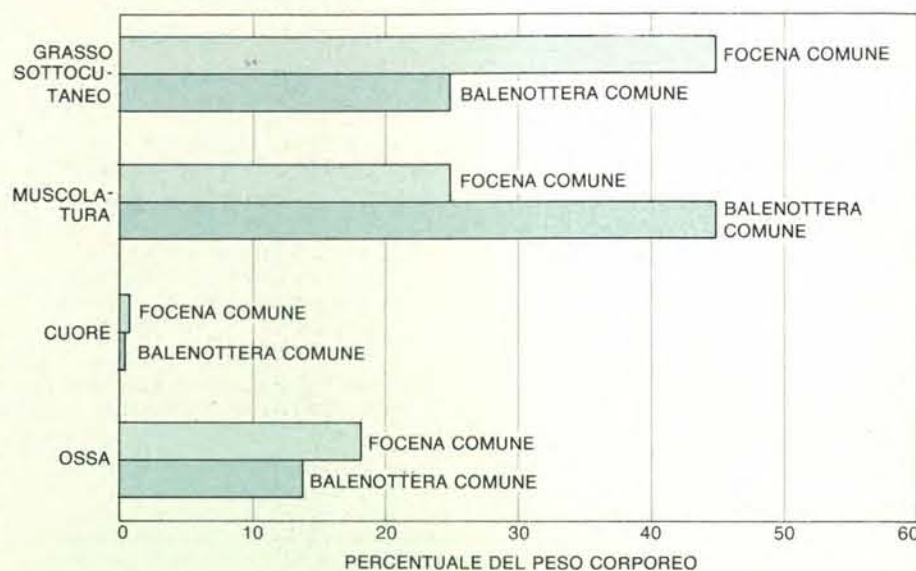
Abbiamo poi tentato di provocare emboli in un altro tursiope addestrato, comandandogli ripetutamente d'immergersi. Non appena la sua respirazione in superficie segnalava che stava recuperando dall'immersione precedente, lo rimandavamo nuovamente in profondità, tanto che nel corso di un'ora rimase al di sotto dei 50 metri per oltre 20 minuti. Comunque, esso non dava segni evidenti di disturbo. Successivamente, con l'aiuto di Robert S. Howard della Scripps Institution, abbiamo misurato la quantità reale di azoto accumulatasi nei muscoli dell'animale a seguito di queste molteplici e ripetute immersioni. Non è stata una cosa semplice: innanzitutto, al termine di questa sequenza di immersioni, si è dovuto istruire il delfino a entrare in un recinto galleggiante, disposto lungo la fiancata della nostra imbarcazione. Questo ci permetteva di far penetrare un sottile ago per prelievo di campioni di gas, attraverso il grasso sottocutaneo, fino alla muscolatura sottostante. Un tubo collegava quest'ago a uno spettrometro di massa (situato su una seconda imbarcazione) che poteva misurare direttamente la concentrazione di azoto presenti nei muscoli. Eseguire queste operazioni in mare, su piccole imbarcazioni, mette veramente a dura prova la pazienza del ricercatore.

Ci attendeva un'altra sorpresa: la concentrazione di azoto disciolto nel tessuto muscolare del tursiope indicava un grado di supersaturazione che sarebbe stato pericoloso per un uomo. Non si poteva far altro che concludere, dunque, che i cetacei sono in grado di tollerare concentrazioni di azoto che, in un nuotatore subacqueo, provocherebbero embolie. Non possiamo ancora spiegare il fenomeno, ma naturalmente è questa la capacità che ci si può attendere da un ordine di mammiferi perfettamente adattato alle grandi profondità.

Quale significato ha avuto l'evoluzione di questa superiore capacità d'immersione per l'ecologia degli oceani? Alcuni animali acquatici, come ad esempio la lontra marina e il manato, si sono evoluti in modo da incidere solo marginalmente sulla catena alimentare degli oceani; altri, come i pinnipedi (foche, leoni marini e trichechi), un sott'ordine di carnivori, si trovano a proprio agio in mare aperto, ma devono raggiungere la terraferma per l'accoppiamento e la riproduzione. Solo i cetacei hanno sviluppato un modo di vita totalmente marino, col vantaggio di un'ampia disponibilità di cibo, sotto forma di altri esseri viventi in mare. Al culmine della catena alimentare marina i misticeti e gli odontoceti si sono affermati come predatori.

Il successo di questi predatori marini fisiologicamente adattati può essere giudicato in base a considerazioni numeriche. Si crede, in generale, che molte specie di cetacei siano in imminente pericolo di estinzione; questa conclusione è fondata in massima parte sulla scarsa conoscenza dei loro attuali livelli di popolazione. Per esempio, il numero totale di balenottere azzurre è maggiore di almeno un ordine di grandezza del numero totale di individui di specie terrestri minacciate quali l'orango e il panda maggiore. In realtà, per la maggior parte, le specie di cetacei sono presenti in quantità consistenti e alcune popolazioni hanno raggiunto un livello tale che servono a scopi turistici: lungo la costa californiana, l'osservazione delle balene è un'attrazione invernale in pieno sviluppo e, nell'Atlantico, le imbarcazioni da diporto che salpano dalla zona di Boston garantiscono ai turisti frequenti occasioni per osservare le balene dopo un'ora o due di navigazione.

Il numero dei capodogli era probabilmente doppio rispetto a quello di oggi, prima dell'inizio della moderna pesca baleniera; comunque il numero attuale è tutt'altro che basso. L'esperto britannico Malcolm R. Clarke valuta l'odierna popolazione di capodogli a circa un milione di esemplari (altri sostengono che sia di 1 milione e mezzo) e ritiene che consumino attualmente 100 milioni di tonnellate di calamari d'alto mare, una risorsa scarsamente attaccata da altri predatori. È istruttivo confrontare la predazione di questa singola specie di cetacei con il pescato oceanico totale in un anno: da 60 a 70 milioni di tonnellate.

Michael E. Q. Pilson dell'Università di Rhode Island ha calcolato il consumo annuale dei mammiferi acquatici che seguono immediatamente nella graduatoria: i pinnipedi. La sua valutazione ha il vantaggio di offrire cifre più attendibili, poiché i pinnipedi devono raggiungere la terraferma per la riproduzione e possono essere contati più accuratamente delle 70 specie di cetacei. Nella maggior parte delle zone di riproduzione di questi animali, sono stati condotti dei censimenti e il loro fabbisogno alimentare annuale è noto in base a quello riscontrato nelle specie in cattività. Pilson ritiene che foche, leoni marini e trichechi consumino 60 milioni di tonnellate annue di pesce: all'incirca il prodotto dell'industria della pesca. Estrapolando dal numero di odontoceti pescati annualmente nel Pacifico orientale, nel corso di operazioni per la cattura dei tonni, il patrimonio mondiale di focene e di delfini può essere calcolato attorno ai 100 milioni di capi. Anche se la popolazione fosse solamente la metà, il suo consumo annuale dovrebbe ammontare in

Nel grafico sono rappresentate le varie percentuali rispetto al peso totale di alcune componenti corporee. Un piccolo odontoceto, la focena comune, è messo a confronto con un grosso misticeto, la balenottera comune delle fredde acque polari. Solo un quarto circa del peso della focena è formato da muscoli, mentre quasi la metà corrisponde al grasso sottocutaneo. La geometria dei rapporti superficie/volume fa sì che la balenottera comune disponga di una spessa guaina di grasso sottocutaneo, corrispondente a un quarto del suo peso totale; quasi la metà è costituita da tessuti muscolari.

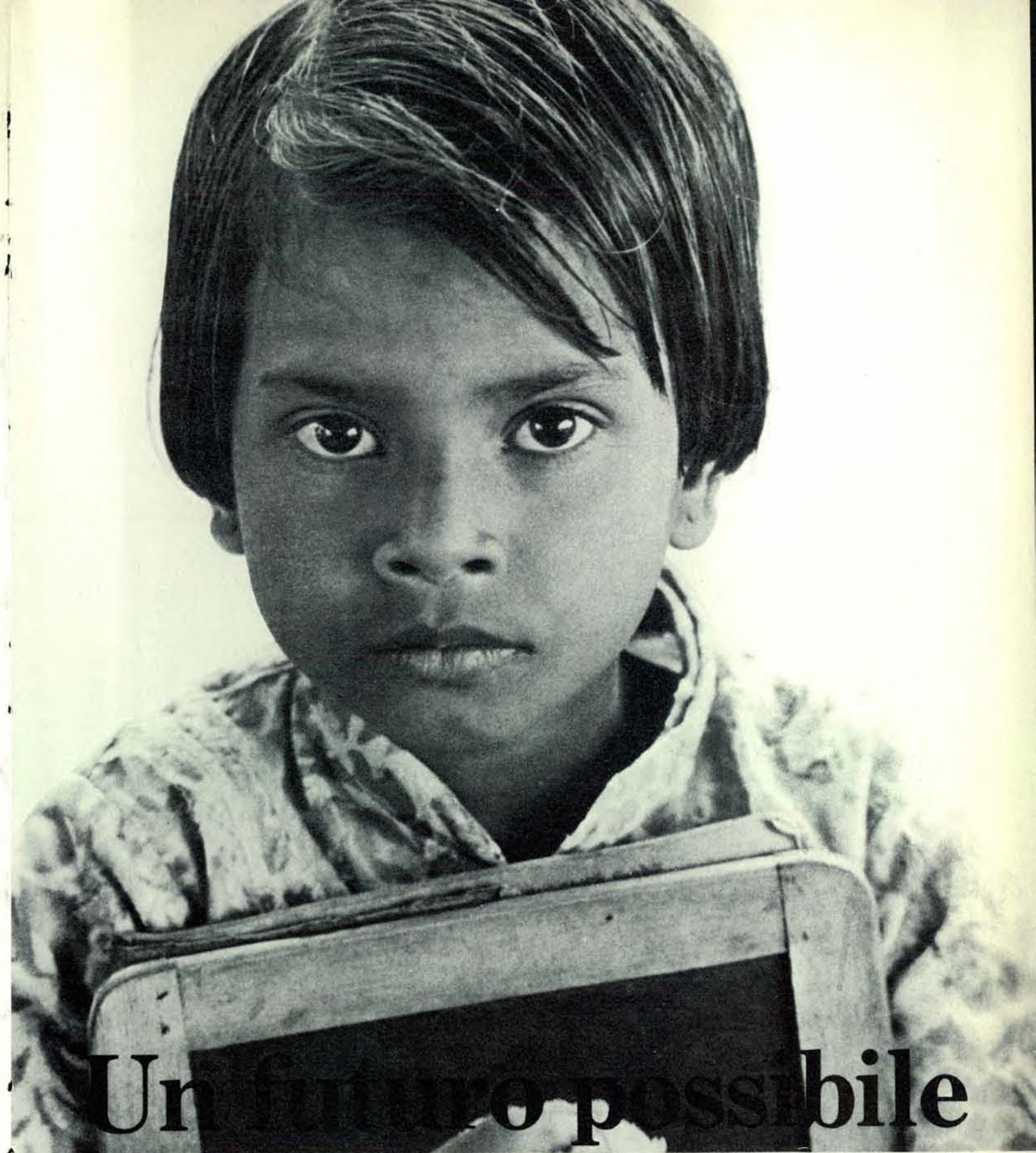
Il ruolo ecologico dei mammiferi marini è altrettanto significativo di quello dei mammiferi terrestri. Il grafico mette a confronto la stima della quantità di organismi marini catturati da mammiferi marini con il valore del pescato totale in tutti i mari del mondo. Un milione di capodogli consumano circa 100 milioni di tonnellate di calamari ogni anno (di più se come sembra i capodogli sono effettivamente in numero maggiore). Da 50 a 100 milioni di focene e di altri piccoli odontoceti consumano da 100 a 300 milioni di tonnellate di pesce. Le stime dei consumi dei grossi cetacei, escludendo i capodogli, sono meno attendibili, mentre foche, leoni marini e trichechi consumano annualmente 60 milioni di tonnellate di pesce. Il pescato totale in tutti i mari del mondo è, invece, compreso fra 60 e 70 milioni di tonnellate *(in colore)*.

effetti a circa 150 milioni di tonnellate.

Cifre così imponenti potrebbero destare sorpresa ma ci affrettiamo a precisare che esse sono puramente ipotetiche. Infatti, salvo quando gli animali oceanici vengono pescati accidentalmente o deliberatamente dall'uomo, è difficile determinare con precisione la consistenza delle loro popolazioni. Due esempi basteranno. Fino a tempi recenti, il delfino di Fraser era noto unicamente per un singolo scheletro trovato nel 1954; ora che la specie è conosciuta, questo delfino è stato osservato in gruppi di centinaia di individui nelle acque delle Filippine e ci si è resi conto che esso è ampiamente distribuito anche nelle zone tropicali degli oceani Indiano e Pacifico. La popolazione di focene di Dall, nel Pacifico settentrionale, era valutata a solo poche decine di migliaia di esemplari finché risultò che le flottiglie giapponesi ne pescavano diecimila all'anno; recenti indagini hanno fatto salire questo valore a un milione e oltre di capi.

È quindi evidente che il ruolo dei cetacei nella catena alimentare marina è ben più significativo di quello dell'uomo. Questo non vuol dire che la pesca eccessiva condotta dall'uomo non abbia influito pesantemente sulle riserve locali, come è accaduto per alcune specie di balene, specie nell'Antartico. Sta però a indicare che, poiché la cattura è in gran parte limitata a poche zone altamente produttive, probabilmente non ha inciso in modo serio sulla dinamica nutrizionale lorda della maggioranza delle zone oceaniche.

Un aspetto della dinamica nutrizionale lorda riguarda il sollevamento di materiale nutritivo proveniente dalle zone profonde, provocato dalle correnti ascensionali al largo dell'Africa Meridionale nell'Atlantico e dell'America Meridionale nel Pacifico. I cetacei adempiono a una funzione analoga su una scala sorprendentemente vasta. Tutte le altre forme di vita marina tendono a muovere il materiale nutritivo verso il basso; i mammiferi marini, legati al regime di superficie in virtù delle loro esigenze respiratorie, provocano il fenomeno inverso; persino il prodotto fecale delle balene non si dirige verso il basso: essendo liquido tende a disperdersi più che ad affondare quando viene emesso.

Per coloro che sono interessati, per ragioni sia emotive sia biologiche, al benessere delle balene e dei cetacei in generale vi sono nuove ragioni di speranza. La caccia a questi animali in mari distanti è un'attività a forte dispendio di energia. Dato che è sempre più difficile trovare le specie più richieste ai fini commerciali, i costi delle campagne di pesca, specie per quanto riguarda il consumo di carburante, sono divenuti talmente elevati da rendere sempre meno redditizia quest'attività. L'ironia della sorte vuole che l'OPEC, l'Organizzazione dei paesi esportatori di petrolio, finirà per produrre, ai fini dell'allentamento della pressione predatoria su questi notevoli mammiferi marini, migliori risultati dei conservazionisti.

La chiusura lampo

Allo sviluppo di questo onnipresente dispositivo, che si affermò una sessantina di anni fa, ha contribuito una miriade di inventori; la sua evoluzione non è ancora giunta al termine

di Lewis Weiner

Il dispositivo di chiusura a cursore noto con il nome di chiusura, o cerniera, lampo, compare su un'ampia varietà di oggetti, che vanno dalle valige ai pantaloni e ai vestiti da donna; per qualunque abitante del mondo occidentale sarebbe inconsueto passare tutta una giornata senza vedersene uno davanti. Ciononostante, questo meccanismo comune e insolitamente sicuro non gode di molta rinomanza. Tutt'altro. Per pura curiosità ho consultato sette enciclopedie che vanno per la maggiore, di cui cinque pubblicate negli Stati Uniti, una in Inghilterra e una in Germania. Nella loro edizione più recente alcune hanno una voce di qualche riga sull'argomento, ma nessuna ha qualcosa che si possa definire un articolo in piena regola. L'origine e lo sviluppo della chiusura lampo sono faccende di cui pochi sanno qualcosa, perfino nelle ditte produttrici. Avendo a che fare da molti anni con questa industria, in parte come inventore di macchine per la fabbricazione di questi congegni e in parte come consulente tecnico per i produttori, mi sono sentito in dovere di presentare una storia completa di questo dispositivo semplice e onnipresente.

Nel 1943 Frank B. Jewett, direttore dei Bell Telophone Laboratories e presidente della National Academy of Sciences, tenne alla New York University una conferenza su «Le promesse della tecnologia». Sottolineando il fatto che «le vere idee creative» hanno origine con gli individui e che nessuno può dire in anticipo quali saranno le idee o dove esse potrebbero nascere, egli disse: «Come esempio dell'impossibilità di prevedere il futuro, anche per le cose semplici, ho pensato spesso alla bassissima probabilità che un qualunque uomo o gruppo di uomini, ove si eccettui quello che ebbe l'idea, progettasse o inventasse la chiusura lampo.»

L'uomo che ebbe realmente l'idea fu Elias Howe, molto più noto per il suo contributo all'invenzione della macchina per cucire. Nel 1851 egli ottenne un brevetto negli Stati Uniti per una chiusura continua e automatica per capi di vestiario. Descrivendola nel suo brevetto, Howe scriveva: «La mia invenzione consiste in una serie di sistemi di aggancio uniti da un cordone di connessione che corre o scorre su nervature.» Il meccanismo di Howe anticipava molte delle caratteristiche delle chiusure lampo dei nostri giorni, ma per ragioni che non sono chiare egli non lo commercializzò mai.

Pertanto l'uomo a cui di solito viene attribuita l'invenzione della chiusura lampo è Whitcomb L. Judson, che nel 1893 ottenne due brevetti negli Stati Uniti per un dispositivo di chiusura automatica fatto funzionare da un meccanismo a scorrimento. Non è dato sapere se Judson avesse una visione limitata delle applicazioni che forse si sarebbero potute trovare per il suo dispositivo di chiusura, o se volesse semplicemente indicare un'applicazione specifica al solo scopo di ottenere il brevetto, ma il titolo del primo dei due brevetti era «Clasp Locker or Unlocker for Shoes» (dispositivo di aggancio per allacciare o slacciare le scarpe). Questi dispositivi di chiusura differivano dalle chiusure lampo dei nostri giorni soprattutto perché il loro movimento era perpendicolare all'apertura che si doveva chiudere, come nei fermagli delle soprascarpe di gomma, prima dell'applicazione della chiusura lampo. Lo schema di Judson consisteva in una serie di ganci separati, ognuno dei quali a sua volta era costituito da due parti collegantisi fra loro, fissate ciascuna a un lato dell'apertura che si voleva chiudere. I ganci potevano essere chiusi a mano, ma Judson forniva anche un congegno scorrevole per chiuderli e aprirli l'uno dopo l'altro.

Nel 1894 Judson e Lewis A. Walker organizzarono la Universal Fastener Company per sfruttare i brevetti di Judson. Walker era un avvocato con una spiccata attitudine per l'organizzazione aziendale. La ditta incominciò nel 1896 a dare impulso a una forma di dispositivo di chiusura detto Universal. Le vendite furono modeste, in parte perché i dispositivi avevano la tendenza ad aprirsi sotto tensione e in parte perché erano piuttosto taglienti e a volte strappavano il tessuto con il quale venivano a contatto.

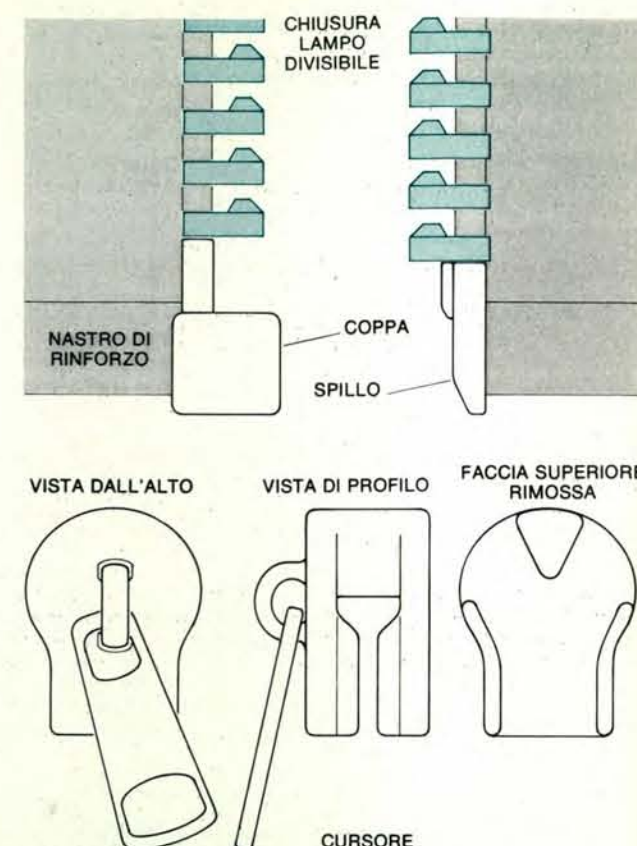
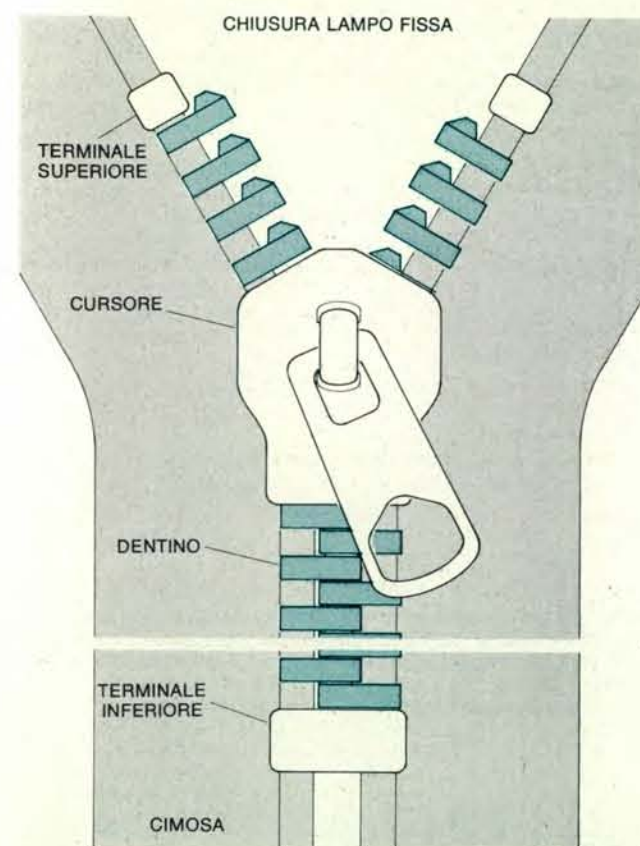
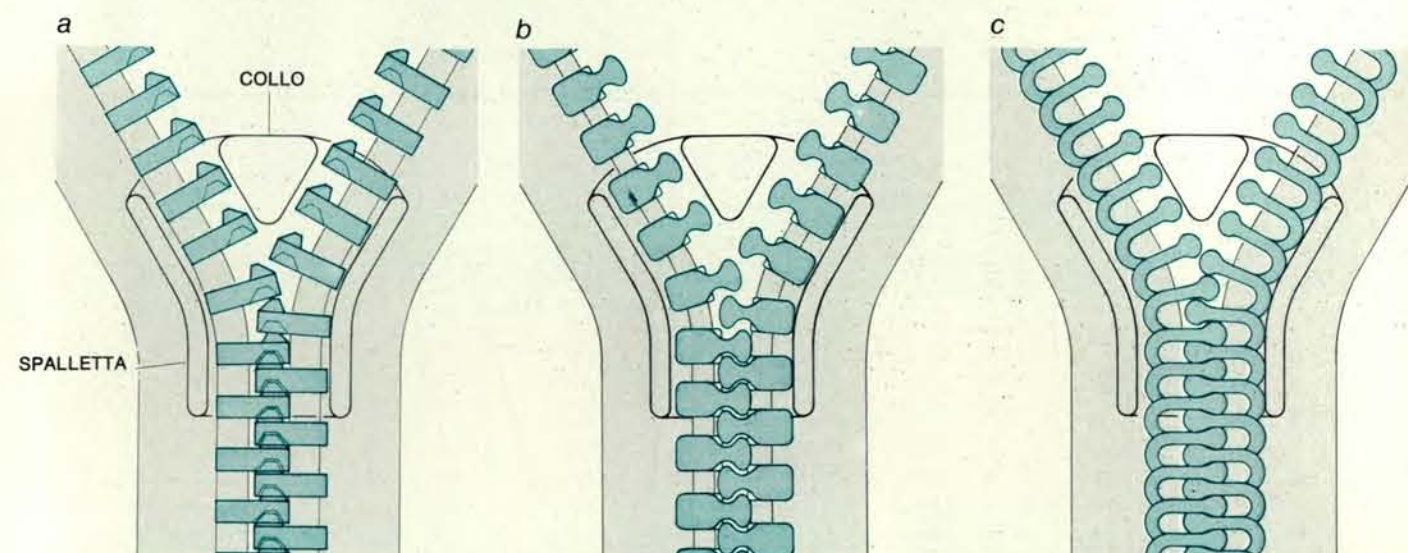
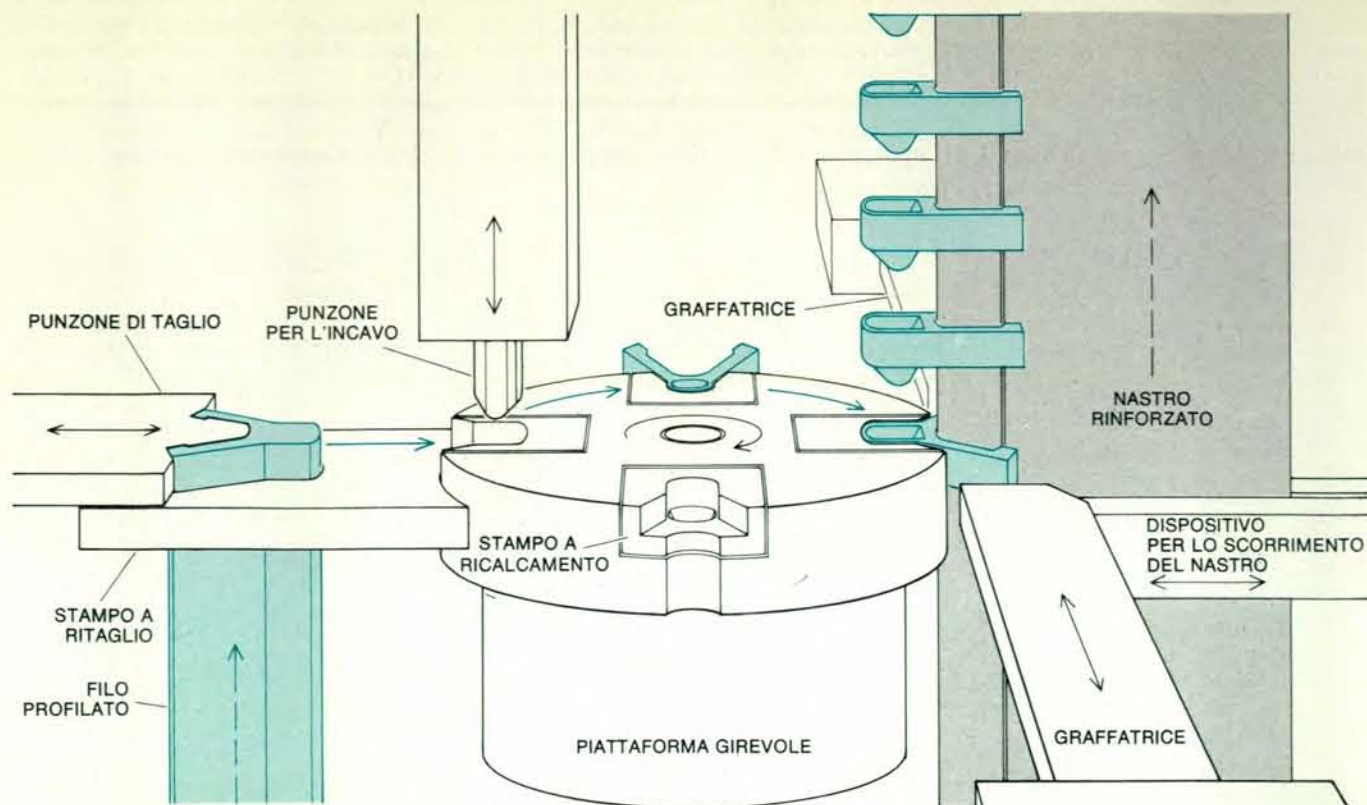
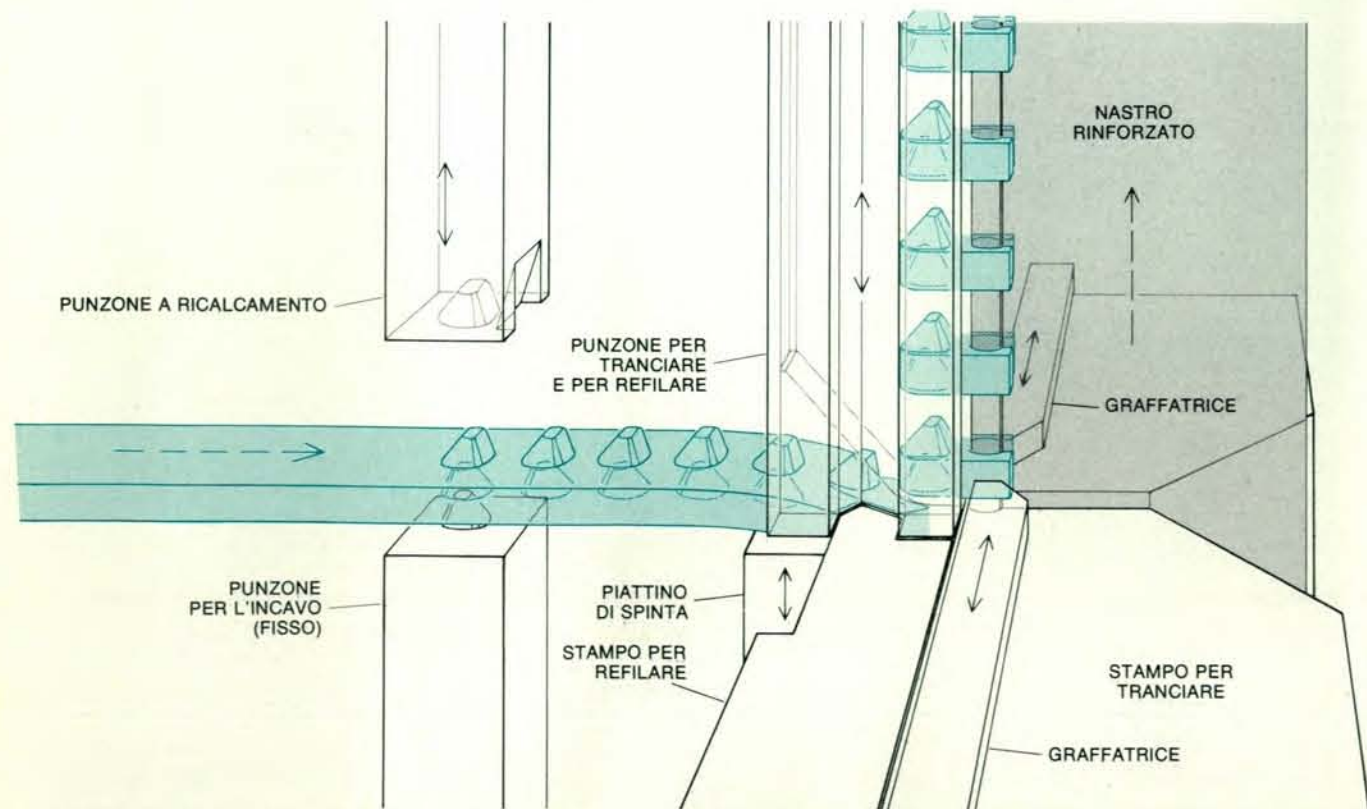
Nel 1904 la società fu riorganizzata sotto la ragione sociale Automatic Hook and Eye Company e offrì un dispositivo di chiusura migliorato, a cui era stato dato il nome C-Curity [che in inglese si legge come la parola *security*, sicurezza, n.d.t.]. Uno slogan pubblicitario diceva «Si tira ed è tutto fatto». Per fabbricare questo dispositivo, che in precedenza veniva costruito a mano, la società poteva disporre di una macchina che Judson aveva brevettato nel 1902. E nel 1906 ebbe l'ulteriore vantaggio di potersi valere dell'opera di Gideon Sundback, un ingegnere che nel corso degli anni migliorò in maniera notevole la chiusura lampo.

Contemporaneamente anche altri inventori andavano via via apportando vari miglioramenti. Nell'anno 1911, Catharina Kuhn-Moos e il suo socio Henri Foster ottennero un brevetto svizzero per un dispositivo che assomigliava molto a una chiusura lampo metallica dei nostri giorni perché non aveva ganci. A quanto pare questi inventori erano troppo avanti rispetto ai loro tempi: il loro dispositivo di chiusura infatti non ebbe successo sul piano commerciale. Nel 1913 a P. A. Aronson, sovrintendente dell'Automatic Hook and Eye, fu concesso un brevetto che presentava il principio della moderna chiusura lampo separabile: un dispositivo che si apriva da tutte e due le estremità.

Le vendite però continuarono a essere modeste, e la ditta arrivò sull'orlo della bancarotta, riuscendo a sopravvivere solo mettendosi a produrre articoli diversi dai dispositivi di chiusura. Poi, nel 1917, Sundback ottenne un brevetto per un dispositivo di chiusura metallico che aveva tutte le caratteristiche della chiusura lampo di metallo dei nostri giorni. Come Sundback ebbe a dire in un'intervista: «La grande necessità era quella di eliminare i ganci». L'invenzione fu tanto importante per la società che quest'ultima cambiò la propria ragione sociale in Hookless Fastener Company. Walker raccolse altri capitali e rifondò l'azienda eleggendosi presidente, carica che conservò fino alla morte.

A poco a poco gli affari migliorarono. Il primo articolo munito di dispositivo di chiusura con meccanismo di scorrimento, fabbricato in grande quantità, fu una cintura con tasche per trasportare denaro, ideata nel 1917 da un sarto di New York: l'articolo godette di molta popolarità fra i marinai della prima guerra mondiale. Quasi tutti i 24 000 dispositivi di chiusura venduti quell'anno dalla società erano proprio per tali cinture. Nel 1918 i dispositivi furono incorporati in circa 10 000 tute fatte per la marina.

Poco dopo la guerra l'introduzione del dispositivo di chiusura con meccanismo di scorrimento sui guanti e sulle borse del tabacco contribuì a rendere popolare questo tipo di chiusura per usi civili. Forse però l'impulso maggiore venne dall'introduzione, nel 1923, del dispositivo in questione sulle soprascarpe di gomma da parte della B. F. Goodrich Company, la quale coniò anche il nome di «zipper», tuttora vivo nella lingua inglese, e lo registrò come marchio di fabbrica per la linea Goodrich di calzature di gomma.

Da quel momento l'espansione fu rapi-

Sono raffigurate qui le varie parti di una chiusura lampo fissa, del tipo usato per i pantaloni, e di una chiusura lampo divisibile, come quelle per le giacche a vento. In inglese i dentini sono chiamati a volte *scoop* perché, visti di profilo, quelli di metallo assomigliano a un mestolo. Il nastro di tessuto serve non solo per tenere fermi i dentini, ma anche per cucire la lampo sugli indumenti. Un nastro munito di dentini o di spirali è detto semicatena. Due semicatene unite formano una catena. Infine, una catena fornita di cursore e di terminali è una chiusura lampo.

Il funzionamento di una chiusura lampo viene illustrato qui come apparirebbe se venisse rimossa la faccia superiore del cursore. Nelle lampo con dentini di metallo *(a)* ogni dentino presenta una sporgenza in alto e un incavo in basso. Via via che il cursore viene sollevato, il suo collo riesce a mettere in posizione i dentini all'angolazione e alla distanza giuste perché la sporgenza di uno di essi si innesti nell'incavo del dentino opposto; le spallette spingono insieme i dentini. Nell'aprire la lampo il collo disinnesta i dentini. Le spirali o i dentini di materiale sintetico *(b, c)* presentano molteplici e svariate forme, ma il principio di apertura e di chiusura della lampo rimane sempre lo stesso.

Una semicatena di metallo viene fabbricata in parecchi modi, uno dei quali è illustrato qui. Un filo profilato a forma di *Y (a sinistra)* viene via via spostato in su nella macchina e un punzone di taglio separa quelli che diventeranno i singoli dentini. I pezzi tagliati sono spinti a mano a mano in uno stampo a ricalcamento che ruota in senso orario. Qui un apposito punzone forma la sporgenza e l'incavo. Dopo due giri di 90 gradi dello stampo il pezzo raggiunge il nastro, che viene spinto fra le gambe della *Y*, le quali vengono strette sul nastro con il bordo rinforzato dalla graffatrice, dopo di che il nastro avanza. Il bordo rinforzato, che non compare su tutti i nastri, offre un altro appiglio alle gambe del dentino.

Un procedimento alternativo per produrre le chiusure lampo di metallo opera con una piattina fornita alla macchina da una coppia di cilindri rotanti. Un punzone a ricalcamento lavora contro un punzone fisso per formare le sporgenze e gli incavi. Una tranciatrice pratica dei dentelli in modo da formare la *Y* di base e nel contempo stacca il dentino anteriore. Prima ancora che questo sia completamente staccato dalla striscia le graffatrici ne piegano le gambe intorno al cordonetto del nastro. Poi la serie di dentini avanza. La macchina può fare 50 dentini al secondo.

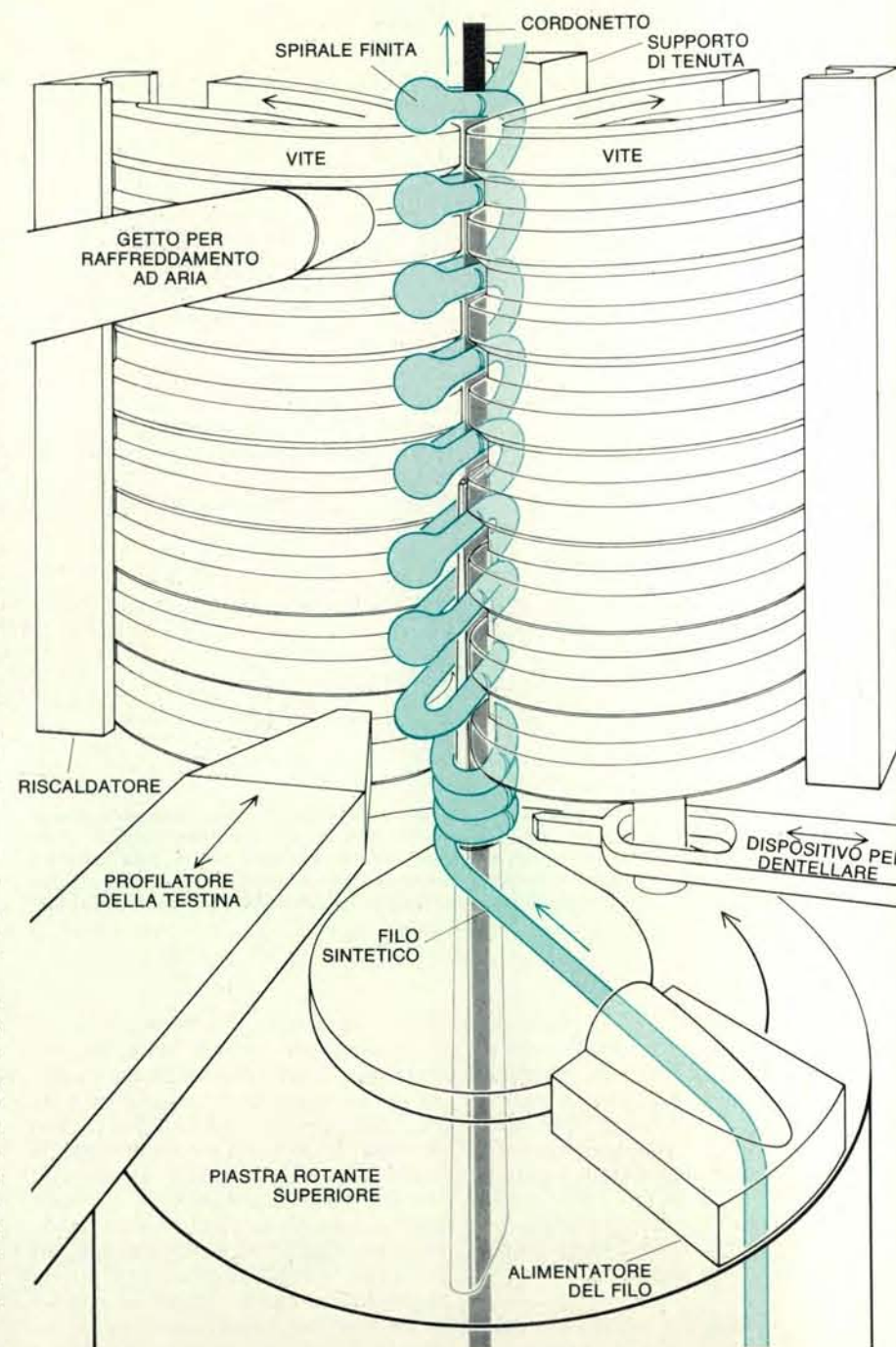
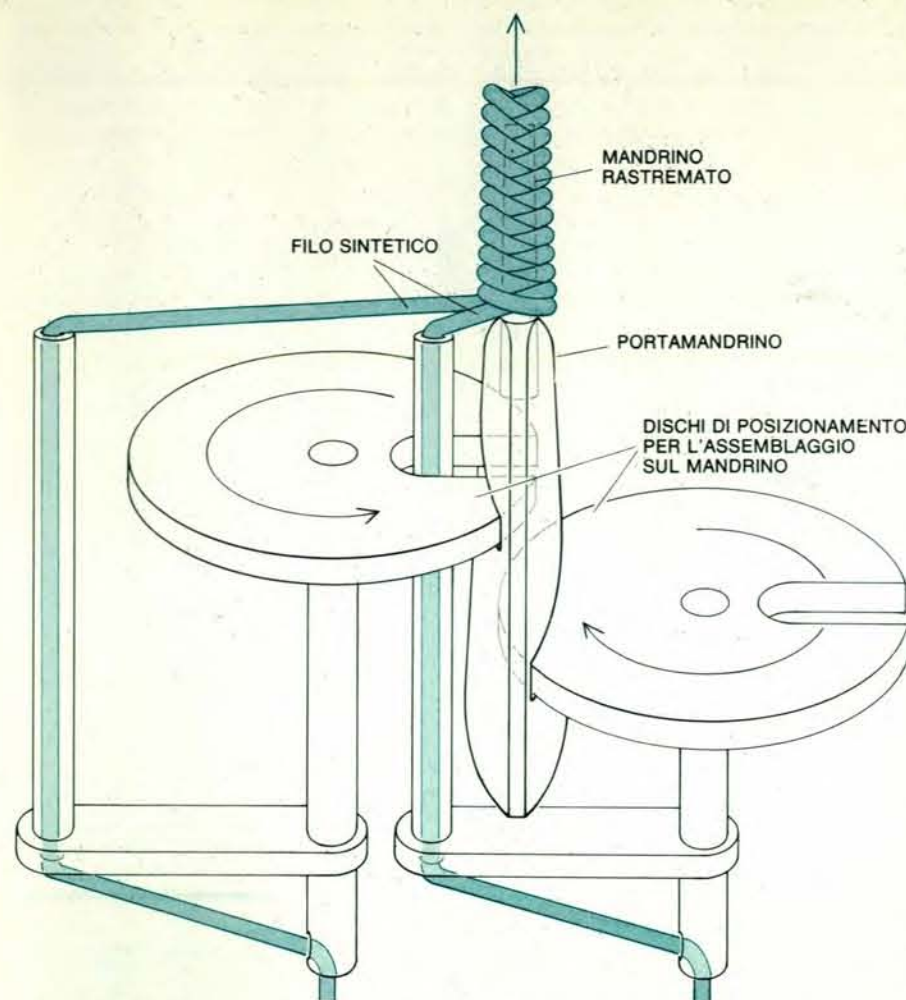
da. La Hookless Fastener, che fu l'unica produttrice di chiusure lampo negli Stati Uniti dal 1917 al 1926, vide salire le proprie vendite da 24 000 pezzi nel 1917 a più di 60 milioni nel 1934. A quel tempo la società aveva mutato in Talon il nome dei propri dispositivi di chiusura e nel 1937 mutò in Talon, Inc. anche la propria ragione sociale. La produzione statunitense di chiusure lampo raggiunse la punta massima di 2,3 miliardi di unità una decina di anni or sono, dopo di che è via via calata per la concorrenza straniera fino a stabilizzarsi nel 1981 intorno a una cifra sotto 1,8 miliardi.

I componenti di fondo di una chiusura lampo di metallo sono i dentini, il nastro, i terminali e il cursore. I dentini sono le unità singole, disposte su un pezzo di nastro. Il nastro munito di dentini è detto semicatena. Due semicatene unite fianco a fianco formano una catena. I terminali sono quelle parti che impediscono al cursore di uscire dalla catena. La catena viene chiusa e aperta dal cursore.

Se si guarda attentamente un dentino di metallo di una chiusura lampo si vedrà una piccolissima sporgenza sulla parte superiore dell'estremità esterna e, in corrispondenza, un incavo su quella inferiore. Insieme, l'una e l'altro danno all'unità una forma che fa pensare a un mestolo, tanto che spesso i dentini delle chiusure lampo in inglese vengono chiamati proprio mestoli *(scoop)*. La sporgenza, o testina, di un dentino si innesta nell'incavo del dentino che, nella serie del lembo opposto, gli sta di fronte; è questa geometria, che si ripete per tutta la lunghezza della catena, a tenere insieme le due semicatene quando la lampo è chiusa.

Se ora si guarda attentamente il cursore, si vedrà che il suo lato destro e quello sinistro formano una flangia, chiamata spalletta. Se si è in grado di scrutare all'interno del cursore, verso la parte superiore, si vedrà un corpo solido a forma di *V*, chiamato collo. Quando si chiude la lampo, la *V* interna e le spallette, agendo insieme, dividono i dentini delle semicatene in maniera sufficiente a separarli e a mantenerli a un'angolazione tale da permettere all'estremità di ogni dentino di innestarsi fra due dentini del lato opposto. Via via che il cursore si muove verso l'alto, le spallette uniscono, dentino per dentino, le due semicatene. Quando si apre la chiusura lampo le spallette mantengono i dentini a un'angolazione tale da permettere ai dentini contrapposti di disinnestarsi, e la *V*, scendendo, li separa.

In passato le chiusure lampo venivano fatte soltanto in metallo. Dopo la seconda guerra mondiale il perfezionamento dei polimeri e dei metodi per foggiarli diede origine a una molteplicità di chiusure lampo in materiale sintetico e in forma di spirale, di serpentina, di meandro e via dicendo. In alcune di esse le spirali contrapposte sono tenute unite da geometrie diverse da quella basata sulle sporgenze e sugli incavi, ma il principio in base al quale il cursore apre e chiude le due parti della catena è circa lo stesso.

Le chiusure lampo sono fatte in vari modi perché certi tipi servono meglio di altri in determinate applicazioni. Sui vestiti, per esempio, devono essere flessibili e non debbono irritare la pelle, e pertanto la scelta migliore pare sia costituita dalle lampo in nylon, nelle quali le semicatene sono costituite da fini spirali. Sui bagagli pesanti, d'altronde, è necessario che le chiusure lampo siano robuste, in modo da poter resistere a carichi notevoli. Qui la

Un dispositivo di avvolgimento che fa uso di un solo filo a sezione circolare produce una spirale modificata per una semicatena di materiale sintetico da un filo di poliestere o di nylon. Due spirali congiunte vengono cucite contemporaneamente su due nastri. Qui si vede la parte superiore del procedimento, a cominciare dal momento in cui il filo sintetico emerge dall'alimentatore e si avvolge intorno a un mandrino rastremato per l'azione di una piastra rotante. Nel medesimo tempo un cordonetto viene tirato verso l'alto attraverso il centro della spirale: esso contribuisce a fissare la spirale stessa al nastro. Via via che si avvolge intorno al mandrino, la spirale viene dentellata da un lato; anche i dentini contribuiscono ad assicurare la spirale al nastro. Due viti riscaldate, che ruotano in senso contrario l'una all'altra, afferrano il filo di materiale sintetico e lo tendono in modo che la parte anteriore diventi orizzontale e appropriatamente intervallata. La parte anteriore della spirale viene compressa per produrre la testina. La spirale viene riscaldata per un certo tempo a una determinata temperatura, dopo di che viene raffreddata in modo da mantenere la forma voluta. Questo è il procedimento detto termofissaggio. Le unità levogire e destrogire lavorano fianco a fianco per produrre semicatene in grado di accoppiarsi.

Una spiralatrice a due fili avvolge entrambi i fili su un unico mandrino. Essa opera abitualmente con un filo di materiale sintetico profilato, anziché con un filo a sezione circolare. I fili vengono avvolti in senso opposto su un mandrino rastremato. Dopo essere stata tolta dal mandrino e termofissata su un'altra ruota, la spirale sintetica viene intessuta, ogni spira separatamente, nella semicatena. Questo procedimento consente un notevole risparmio di materiale.

cosa migliore è una lampo di grandi dimensioni con dentini di metallo oppure con spirali o denti in materiale sintetico.

Le dimensioni di una catena vengono identificate misurando la larghezza delle due serie di dentini (o delle spirali) a lembi uniti. Per le catene di metallo i numeri vanno da 2 (0,135 pollici) a 7 (0,35 pollici). Per le catene sintetiche i numeri variano da 3 a 9, che corrispondono grosso modo ad altrettanti millimetri.

La catena per chiusura lampo viene prodotta in un unico pezzo continuo e poi tagliata in tante chiusure lampo singole, a ciascuna delle quali vengono aggiunti un cursore e dei terminali che, a seconda del tipo di chiusura, limitano variamente il percorso del cursore. Una chiusura lampo, come quelle che si applicano sulle giacche a vento, ha un terminale nella parte superiore di ogni semicatena e una coppa e uno spillo in quella inferiore. Le chiusure lampo fisse (cioè aperte solo in alto), come quelle che si applicano ai pantaloni, hanno un unico terminale nella parte inferiore della catena e uno nella parte superiore di ciascuna semicatena.

Consideriamo ora la fabbricazione delle chiusure lampo, dividendo queste ultime in due categorie, a seconda del fatto che l'elemento di chiusura sia di metallo o sintetico. Descriverò tre metodi per le catene di metallo. In ciascun caso il lavoro è svolto da una macchina interamente automatica, e un solo addetto può badare a più macchine. Fondamentalmente, nei primi due metodi i singoli dentini vengono prodotti in rapida successione e fissati su un nastro di tessuto che si va via via srotolando. Il bordo del nastro è rinforzato, in modo che le gambe dei dentini vi si stringano saldamente.

Il primo processo ha inizio con un filo da sezione circolare di rame, di alluminio o di una lega di nichel-argento. Il filo è sottoposto a varie operazioni in un laminatoio, che lo profila a forma di Y, sì che, quando viene tagliato a fette come un filone di pane, si ha una serie di singole Y, ognuna delle quali diventa un dentino: i bracci della Y sono le gambe che si fissano al nastro con il bordo rinforzato.

Via via che il filo profilato viene immesso nella macchina che fa le chiusure lampo, una punzonatrice taglia un pezzo profilato a *Y* e lo spinge in uno di quattro stampi a ricalcamento, situati su una piattaforma girevole. Un'altra punzonatrice scende sulla *Y* e forma la sporgenza e l'incavo. Dopo due rotazioni di 90 gradi della piattaforma girevole, durante ognuna delle quali un pezzo grezzo viene aggiunto in un altro stampo, il pezzo grezzo raggiunge il nastro con il bordo rinforzato. Il bordo rinforzato del nastro viene spinto nel dentino, e delle graffatrici chiudono le gambe di quest'ultimo intorno a esso. Il nastro viene poi tirato di un paio di volte lo spessore del dentino più un dieci per cento circa, una distanza calcolata in modo da creare spazio per il dentino successivo e per il dentino gemello del lembo opposto che si innesterà fra i due quando la lampo verrà chiusa.

Questo metodo di fabbricazione per chiusure lampo di metallo ha una sessantina d'anni di età. In effetti, la macchina è una diretta discendente della prima macchina automatica, inventata da Sundback nel 1923, usata per fare le catene. Prima di allora il sistema generale per produrre le chiusure lampo era quello di punzonare i dentini a uno a uno. I dentini venivano burattati per eliminare i bordi taglienti, dopo di che venivano laminati e inseriti a mano in un'apparecchiatura. Quando questa apparecchiatura era piena, un nastro cordonato veniva fatto passare in mezzo alle gambe del dentino, che poi venivano chiuse sul nastro per mezzo di una pressa meccanica. In seguito la procedura manuale fu meccanizzata, ma era ancora lenta e tale da non dare completo affidamento. Inoltre era dispendiosa: la punzonatura dei dentini lasciava circa il 40 per cento di scarti. La macchina automatica di Sundback, invece, non lasciava scarti e quindi ebbe un successo strepitoso. I procedimenti basati su un filo a sezione circolare sono peraltro relativamente lenti e non sono molto usati oggi.

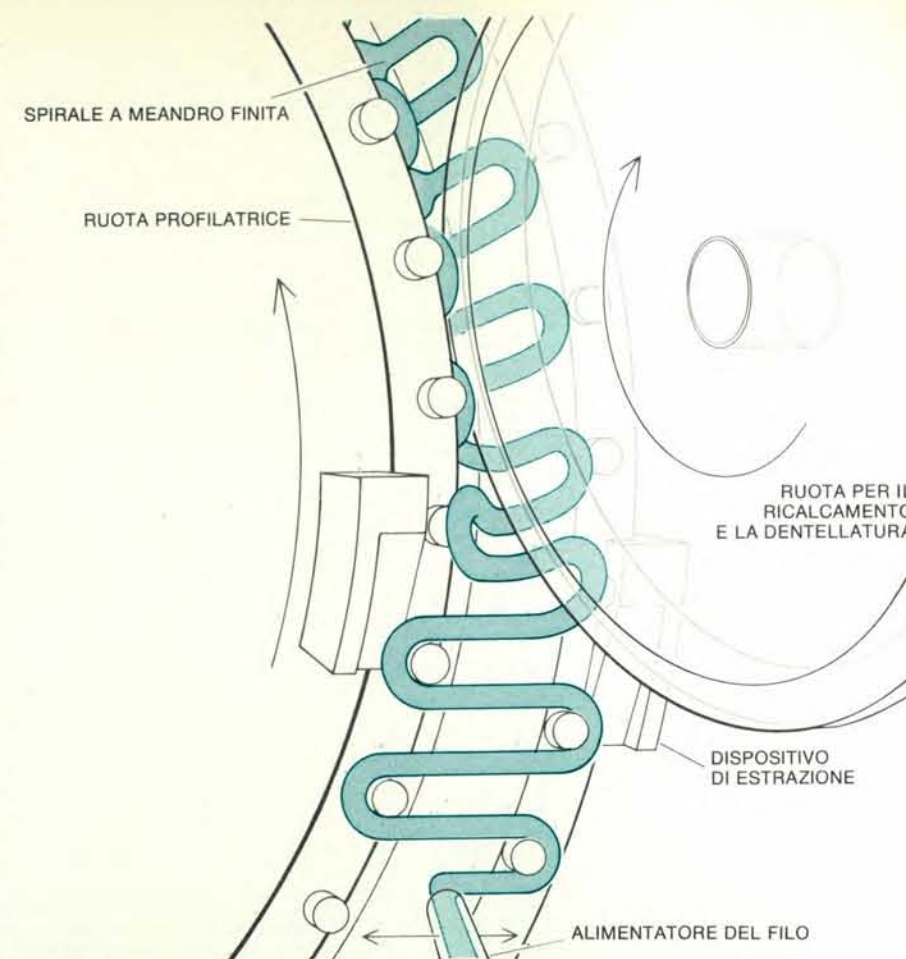
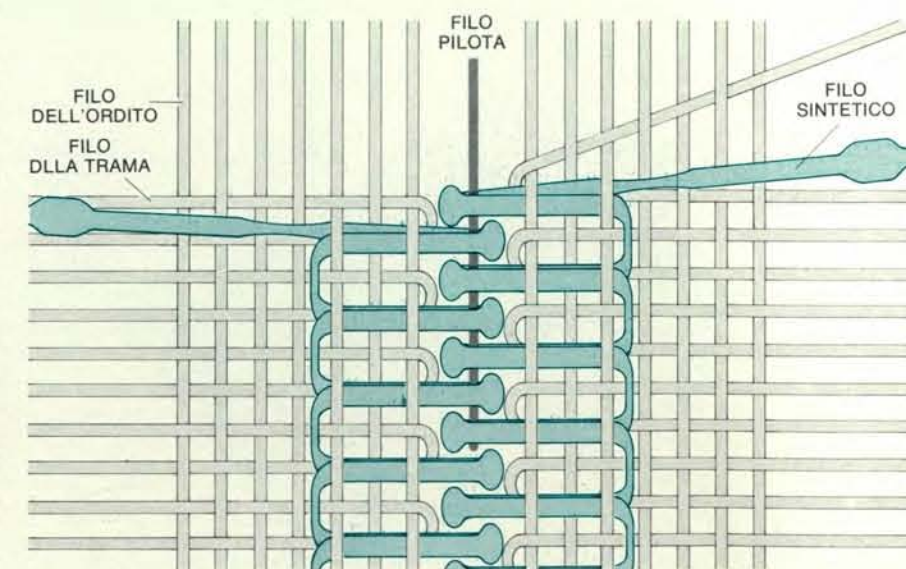
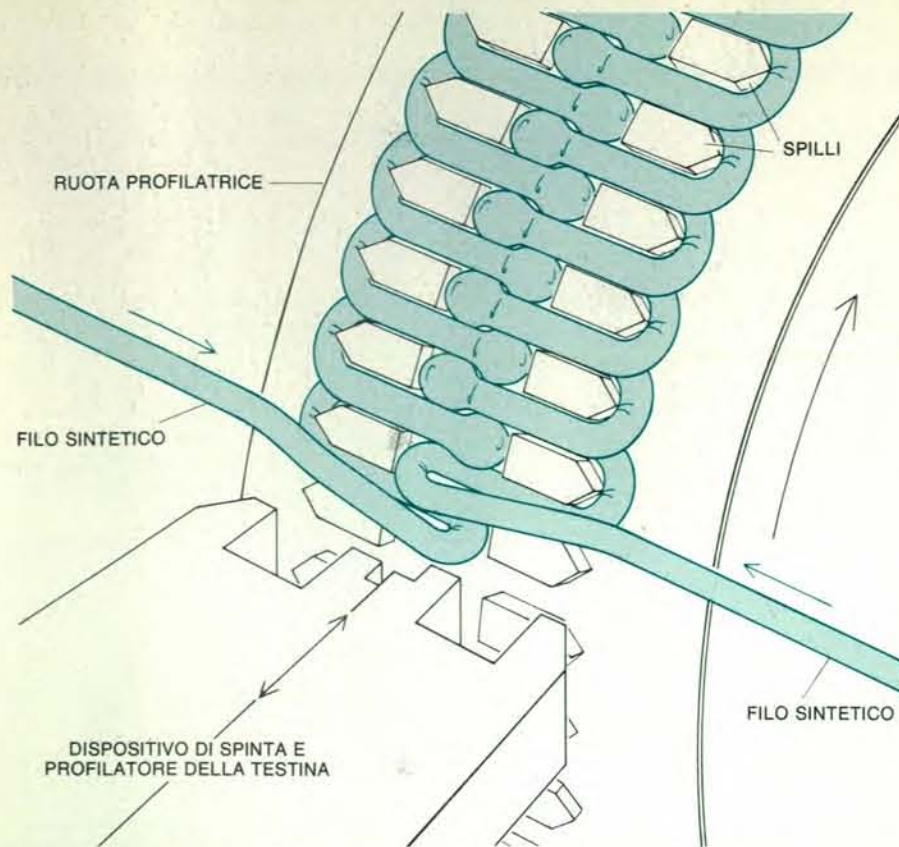
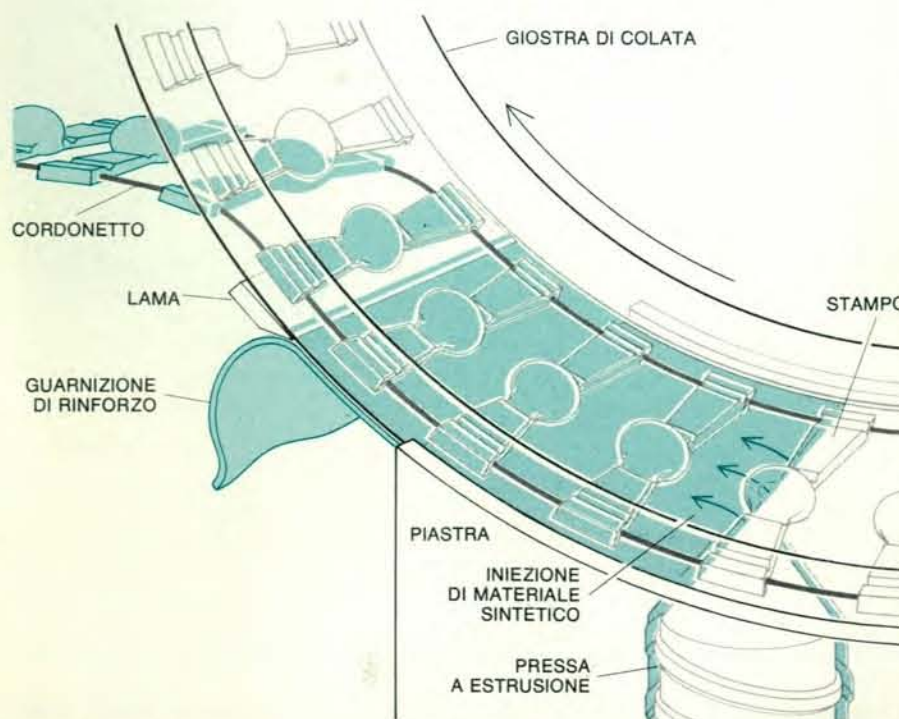
Nel secondo procedimento, che si rifà a un brevetto ottenuto nel 1940 da Frederick Ulrich della Conmar Products Corporation, la macchina lavora non su un filo a sezione circolare ma su una piattina. Due cilindri rotanti forniscono la piattina alla macchina, dove una punzonatrice a ricalcamento lavora contro una punzonatrice fissa in modo da formare la sporgenza e l'incavo su quello che diventerà il dentino singolo della serie continua. Poi una tranciatrice pratica dei dentelli su entrambi i lati del filo e, nel medesimo tempo, stacca il dentino successivo dalla piattaforma. Prima che il pezzo grezzo sia tagliato completamente, le sue gambe vengono piegate intorno al bordo rinforzato del nastro. Una macchina ben progettata di questo genere lavora alla velocità di circa 50 dentini al secondo.

Nel terzo procedimento, la cui origine va ricercata in un brevetto ottenuto nel 1932 da Gustav Johnson, i dentini non vengono stampati, ma pressofusi direttamente sul nastro. Il nastro si muove attraverso uno stampo aperto composto di due parti, che ha una serie di cavità corrispondenti ai dentini. Chiuso lo stampo, vi si inietta una lega di zinco sotto pressione. Lo stampo è raffreddato ad acqua e, quando il metallo fuso solidifica, si apre. Il nastro con i dentini si porta nella posizione successiva e il materiale di scarto viene tagliato via.

Una volta prodotta con uno qualunque di questi tre procedimenti, la semicatena di metallo viene sottoposta ad alcune operazioni di rifinitura. Due semicatene vengono unite in una catena continua da un'attrezzatura di chiusura che assomiglia al cursore di una chiusura lampo standard. La catena chiusa passa attraverso rulli che la comprimono nello spessore voluto, dopo di che passa in mezzo a una serie di spazzole metalliche che eliminano i bordi taglienti. La tappa successiva è un bagno di appretto. Uscita da tale bagno, la catena viene spremuta da appositi rulli e fatta passare sopra cilindri caldi di essiccazione. Lo scopo di questa operazione è quello di stirare il nastro. Alla fine i dentini ricevono un'applicazione di cera fusa, che fa scorrere agevolmente il cursore quando la lampo è nuova. (Quando questa si sarà rodata, il cursore funzionerà senza cera.) La catena continua finita viene avvolta intorno a una bobina ed è pronta per essere montata in modo da formare una chiusura lampo.

La chiusura lampo in materiale sintetico del giorno d'oggi fece la sua prima comparsa in Germania dopo la seconda guerra mondiale. Durante il conflitto, i produttori tedeschi del settore avevano perso fabbriche e macchinari e quel che si era salvato era ormai antiquato. Dovendo ricostruire la propria industria, i tedeschi decisero di sviluppare la chiusura lampo in materiale sintetico.

La società pioniere in questo campo fu la Opti-Werk GMBH della Germania Occidentale. Come sempre, però, lo sviluppo delle nuove chiusure fu opera di molti inventori. Nel 1942 l'americano Alden W. Hanson ottenne un brevetto per un sistema che cuciva una spirale di materiale sintetico su un nastro; nel 1951 Nicholas A. Wahl ne ricevette uno per un metodo che consentiva di avvolgere due fili di plastica su un unico mandrino. Dopo di che, a J. R. Ruhrman e ai suoi colleghi fu concesso un brevetto tedesco per una catena di materiale sintetico a meandro. Ad A. Gerlach e ai suoi colleghi e, automaticamente, alla ditta William Prym-Werke furono concessi brevetti per un filo di materiale sintetico dentellato che poteva essere intessuto in un nastro. Un altro brevetto fu concesso nel 1968 all'australiano E. E. Cuckson e ai suoi colleghi per un sistema che fondeva i dentini in chiusure lampo in continuo. Una spiralatrice che piega due fili a sezione circolare di materiale sintetico in una ruota di termofissaggio fu inventata negli Stati Uniti nel 1969.

Una semicatena sintetica ha spirali di poliestere o nylon, costituite da spirali modificate, meandri o altre forme. Oggi nella maggior parte delle chiusure lampo, ove si eccettui la sola versione in cui i

Il filo di sezione circolare, in materiale sintetico, viene profilato a forma di meandro da una coppia di ruote. L'alimentatore del filo si muove a intermittenza e alternativamente rispetto alla ruota profilatrice, e anche quest'ultima si muove a intermittenza in modo da permettere al filo di stendersi tra gli spilli. Dispositivi di estrazione fissi sollevano il filo da questi ultimi. La seconda ruota comprime su quella profilatrice la configurazione a meandro dandole la forma di una *U*, la plastica viene termofissata e vengono modellate le testine. Alla fine i meandri vengono cuciti sui nastri.

Viene illustrata la tessitura di un filo sintetico profilato in una catena di una chiusura lampo. Due navette portano il filo della trama e due quello sintetico. Il filo pilota è fissato al fusto del telaio ed esce dalla catena via via che la tessitura procede. Le catene così fatte sono della migliore qualità e adatte per indumenti. L'operazione è lenta, perché solo una delle quattro navette lavora mentre le altre aspettano, e così il costo di fabbricazione è elevato. La chiusura tessuta a telaio non è adottata negli Stati Uniti, dove però si importano molte lampo di questo tipo. Ultimamente è stato sviluppato in Italia un sistema brevettato analogo molto rapido, ma su telai ad ago. Questo brevetto è attualmente alla base della produzione della Fastex s.r.l. di Monza.

Questo dispositivo che avvolge due fili sintetici a sezione circolare produce contemporaneamente la spirale destra e quella sinistra. I due fili di materiale sintetico vengono immessi dai due lati nella ruota profilatrice, in uno spazio ristretto dove formano dei cappi. Il dispositivo di spinta e il profilatore della testina spingono il filo tra gli spilli. Dopo la fase di termofissaggio il prodotto viene tolto dalla ruota sotto forma di due spirali chiuse, che poi vengono cucite sui nastri.

La fusione dei dentini sintetici avviene su una giostra di colata che porta delle cavità corrispondenti alla forma dei dentini e disposte a intervalli regolari. I due cordonetti forniti a ogni linea di dentini collegano questi ultimi e rendono più facile il distacco delle singole unità della ruota. Materiale sintetico semifuso viene distribuito dalla pressa a estrusione alle cavità, dove viene tenuta in posizione dalla piastra e raffreddata finché non solidifica. La lama asporta il materiale eccedente. In un'operazione distinta (non illustrata qui) i dentini vengono piegati a forma di *U*.

dentini sono fusi direttamente sul nastro, le spirali di materiale sintetico sono cucite o intessute nel nastro. Tanto il poliestere quanto il nylon hanno un punto di fusione superiore a 204 gradi centigradi, sì che gli indumenti con chiusure lampo sintetiche si possono stirare senza pericolo.

Ai fini di questa discussione descriverò due macchine per la fabbricazione di chiusure lampo sintetiche. Una è la macchina che avvolge un filo di materiale sintetico a sezione circolare intorno a un mandrino. Le sue parti essenziali sono il mandrino fisso rastremato e la sezione non rotante per il cordonetto, l'uno e l'altra montati su un albero disassato. (Il cordonetto corre verticalmente all'interno della spirale della semicatena e contribuisce a fissare la spirale al nastro.) Una bobina rotante fornisce il filo sintetico e il cordonetto è sfilato da una bobina conica. Altre parti importanti sono due viti riscaldate, un dispositivo per intagliare e un altro di formatura delle testine.

Il filo di materiale sintetico esce dalla bobina rotante, passa in mezzo a due piastre rotanti ed entra in un alimentatore che lo avvolge intorno al mandrino. Nel medesimo tempo il cordonetto viene tirato attraverso l'albero disassato e una scanalatura praticata nel mandrino, sì che finisce col passare attraverso il centro della spirale. Non sempre si fa uso del cordonetto. A ogni giro dell'alimentatore la spirale avanza sul mandrino. Via via che procede, viene dentellata sul fianco; i dentelli sono un altro modo per contribuire a fissare le spirali quando vengono cucite sul nastro.

Le due viti riscaldate, che ruotano in senso opposto l'una all'altra in direzione della spirale dentellata, afferrano quest'ultima e la tendono in modo che la sua estremità superiore sia orizzontale e intervallata in modo corretto. Una barra fissa sistemata dietro viti tiene le spirali nella posizione voluta. Le viti vengono riscaldate in modo che stabilizzino il materiale termoplastico della spirale. Dopo essere stata tesa, la parte anteriore della spirale viene compressa in modo da creare le testine. Alla fine, la spirale viene raffreddata con un getto d'aria. Spesso le viti sono sostituite da ingranaggi che servono a tendere e a foggiare la spirale, dopo di che il termofissaggio viene effettuato mediante riscaldamento a induzione.

In una catena del tipo a spirale le due semicatene opposte devono avere delle spirali avvolte in senso contrario l'una all'altra. Sono quindi necessarie due macchine (o una macchina con due dispositivi distinti) per avvolgere una spirale sinistra e una destra. Normalmente le due macchine (o i due dispositivi) operano fianco a fianco in modo che le due spirali si possano unire in una spirale doppia. La spirale doppia così unita è pronta per essere cucita sul nastro.

I dentini di materiale sintetico fusi in continuo vengono prodotti su una giostra di colata che lungo il proprio perimetro ha, a intervalli regolari, cavità corrispondenti alla loro forma. Ogni fila di dentini viene dotata di un paio di cordonetti, i quali tengono insieme le serie di dentini e facilitano la rimozione delle semicatene dalla giostra di colata. Ogni stampo per dentino che passa sulla giostra di colata viene riempito di materiale sintetico semifuso. Una piastra tiene fermo il materiale al suo posto finché non solidifica, dopo di che una lama fissa rimuove dalle semicatene la plastica eccedente. Poi le semicatene passano attraverso una piegatrice, dove vengono piegate nella forma voluta e termofissate. La tappa successiva è quella di cucire le semicatene sul nastro della chiusura lampo.

Una volta fatta, la catena sintetica viene sottoposta a varie operazioni di finitura. Viene messa in un bagno di appretto e fatta passare fra rulli spremitori e in un essiccatore allo scopo di stirare il nastro. Alla fine la catena continua viene avvolta intorno a una bobina di carta.

Per chiudere o aprire una lampo è necessario un cursore. Il corpo del cursore e il tiretto sono tranciati da una striscia di acciaio dolce od ottenuti per pressofusione da zinco. In genere i cursori tranciati vengono messi sulle catene di metallo e quelli pressofusi sulle catene di materiale sintetico. I cursori usati in queste ultime devono avere all'interno del proprio corpo delle nervature per costringere la spirale nella posizione giusta quando la lampo si sta chiudendo. È difficile fare delle nervature in un processo di tranciatura, ed è per questo che sulle catene sintetiche il cursore più comune è quello ottenuto per pressofusione.

Per rendere il cursore inossidabile, le parti componenti vengono placcate in ottone o in zinco. Un'eccezione appare nei cosiddetti cursori automatici. I cursori di questo tipo hanno una molla che li tiene fermi finché non viene sollevato il tiretto; la molla è in genere di acciaio inossidabile. Spesso il cursore è smaltato in modo da intonarsi al colore della catena. Il corpo, il tiretto e (nei cursori automatici) la molla sono montati in una macchina speciale.

L'ultima fase della fabbricazione di una chiusura lampo è quella del montaggio della catena, del cursore (due cursori per il tipo di lampo che si può aprire e chiudere da tutte e due le parti) e dei terminali. Le macchine usate per il montaggio sono diverse per le chiusure lampo di metallo e per quelle sintetiche, ma i metodi sono fondamentalmente gli stessi. Una tipica operazione di montaggio ha inizio con la chiusura lampo in catena continua su una bobina. Una coppia di estrattori circolari srotola la catena, tirandola innanzitutto attraverso un dispositivo che rimuove i dentini a determinati intervalli in modo da creare un tratto di nastro vuoto e che simultaneamente attacca il terminale inferiore; tagliando lo spazio morto all'altezza del suo punto mediano si determina la lunghezza della lampo e si ottiene quel tanto di nastro in più necessario per cucirlo appropriatamente su un indumento. Lo spazio morto si può creare anche con la macchina usata per fare la catena: in questo caso si omettono dei dentini. Dopo di che la catena viene fatta passare attraverso una macchina che introduce il cursore nella catena stessa attraverso lo spazio morto e si applica il terminale superiore. L'ultima fase è costituita dalla pezzatura, cioè dal taglio della catena all'altezza del centro di uno spazio morto. A volte le operazioni avvengono in combinazione. Come risultato si ha una chiusura lampo singola, pronta per essere applicata a un indumento, a una valigia o a qualsiasi altro prodotto per il quale era stata progettata.

Per le chiusure lampo divisibili il procedimento di fabbricazione è leggermente diverso. Qui il nastro di cotone, rivestito di nylon, viene fuso all'altezza dello spazio morto come elemento di rinforzo, dopo di che si applicano i terminali superiori. Il nastro rivestito di nylon viene tagliato lungo un prolungamento dell'asse della chiusura lampo. Il cursore viene fatto scorrere su una sola semicatena e viene montata anche la coppa. Sull'altra semicatena viene montato lo spillo. Quando la lampo è tutta aperta, lo spillo esce dalla coppa attraverso il cursore e la chiusura lampo si separa completamente, con il cursore che rimane sulla propria semicatena, trattenuto dalla coppa.

Si potrebbe supporre che non ci sia molto da attendersi a proposito di nuovi sviluppi in un dispositivo così semplice e standardizzato come una chiusura lampo. In realtà però esistono parecchie possibilità, soprattutto per quelle in plastica. Una riguarda le chiusure lampo molto più grandi o molto più piccole di quelle attualmente in commercio. Quelle più grandi sarebbero destinate a svolgere un lavoro pesante chiudendo, poniamo, dei divisori a tenda che permettono di creare piccoli luoghi d'incontro in un grande spazio. Le chiusure lampo più piccole servirebbero per quei prodotti che rispecchiano l'attuale tendenza verso la miniaturizzazione, come gli astucci in grado di adattarsi a radio o registratori tascabili.

Un'altra prospettiva riguarda le chiusure lampo a buon mercato destinate a indumenti di carta da buttare via dopo l'uso e ad altri oggetti da usare una volta e poi eliminare. In un'applicazione più insolita, è possibile prevedere una chiusura lampo che sia robusta, a tenuta d'aria e al tempo stesso tanto flessibile da poter essere utilizzata sugli indumenti degli astronauti. Se una chiusura lampo di questo genere verrà sviluppata, farebbe probabilmente la sua comparsa su prodotti di consumo che richiedono dispositivi di chiusura con tali caratteristiche.

Un esempio finale è costituito da un'applicazione che è tuttora alla ricerca di una chiusura lampo. In chirurgia esiste la necessità di chiusure lampo a tenuta d'aria e chimicamente inerti da usare al posto dei punti nelle incisioni che potrebbe essere necessario riaprire per accedere a una protesi qual è, per esempio, un *pacemaker*. Parecchi chirurghi hanno espresso il desiderio di un congegno di questo genere; progettarlo e fabbricarlo sarà una sfida per i ricercatori medici e per l'industria delle chiusure lampo.

Georg Cantor e la teoria degli insiemi transfiniti

Quanto è grande un insieme infinito? Cantor ha dimostrato che esiste una gerarchia di infiniti, ciascuno più «grande» del precedente, e oggi la sua teoria degli insiemi è uno dei cardini di tutta la matematica

di Joseph W. Dauben

La natura dell'infinito è sempre stata un tema oggetto di controversia. Compare, nell'antichità, nei famosi paradossi di Zenone di Elea, il quale sostenne con chiarezza disarmante che il movimento è impossibile perché altrimenti un oggetto dovrebbe passare per un numero infinito di punti in un tempo finito. Il successo della fisica newtoniana è dovuto in gran parte all'introduzione da parte di Newton del calcolo di velocità infinitesime di variazione, anche se per oltre 200 anni nessuno ha saputo dare una formulazione matematicamente rigorosa della sua strana, ma efficace, idea. In epoca moderna problemi attinenti l'infinito sono apparsi nella teoria astratta degli insiemi, teoria che fornisce un fondamento per praticamente tutta la matematica contemporanea. Inoltre, spesso, in tutta la sua storia, l'idea di infinito ha assunto toni teologici che hanno avuto una parte non secondaria nell'accettazione o nel rifiuto del concetto e delle dottrine matematiche e filosofiche che lo accompagnavano. Tutte queste correnti di pensiero convergono nella vita e nelle opere del matematico tedesco Georg Cantor.

La sostanza del lavoro che occupò tutta la vita di Cantor è ben nota: sviluppando quella che chiamava l'aritmetica dei numeri transfiniti diede un contenuto matematico all'idea di infinito attuale. Nel contempo pose le fondamenta della teoria astratta degli insiemi e fornì contributi significativi alla fondazione del calcolo e all'analisi del continuo dei numeri reali. Il risultato più notevole di Cantor fu la dimostrazione, in forma matematicamente rigorosa, del fatto che quello di infinito non è un concetto indifferenziato. Non tutti gli insiemi infiniti hanno le stesse dimensioni: pertanto, è possibile confrontare fra loro insiemi infiniti. L'insieme dei punti su una retta, per esempio, e l'insieme di tutte le frazioni sono insiemi infiniti, ma Cantor riuscì a dimostrare, in un senso ben preciso, che il primo insieme è più grande del secondo. Le idee di Cantor erano così sconvolgenti e così controintuitive per i suoi contemporanei, che il grande matematico francese Henri Poincaré condannò la teoria dei numeri transfiniti come «una malattia» da cui un giorno la matematica sarebbe stata curata. Leopold Kronecker, uno dei maestri di Cantor e grande esponente dell'*«establishment»* matematico tedesco, arrivò al punto di attaccare Cantor personalmente, definendolo addirittura un «ciarlatano scientifico», un «rinnegato» e un «corruttore della gioventù».

È ben noto, anche, che Cantor soffrì per tutta la vita di una serie di «esaurimenti nervosi», che divennero sempre più frequenti e sempre più debilitanti con l'età. Questi esaurimenti erano probabilmente i sintomi di una malattia mentale organica; uno studio recente condotto dallo storico della matematica inglese Ivor Grattan-Guinnes, sulla base della valutazione della documentazione medica di Cantor da parte di psicologi della Halle Nervenklinik (l'ospedale per malattie mentali di Halle, nella Germania Orientale), lascia pensare che Cantor fosse vittima di una psicosi maniaco-depressiva. Ciononostante, i primi biografi di Cantor hanno avuto buon gioco a presentare questo matematico, che cercava di difendere la sua complessa teoria ma soffriva per periodi sempre più lunghi di esaurimenti mentali, come la vittima sfortunata delle persecuzioni dei suoi contemporanei.

Questi racconti distorcono la verità, banalizzando le genuine preoccupazioni intellettuali che motivano alcune delle forme più serie di opposizione alla teoria di Cantor da parte dei contemporanei; e inoltre non danno il giusto credito alla potenza e all'ampiezza della difesa che Cantor mise in atto per le proprie idee. Inizialmente egli respinse l'idea dei numeri transfiniti, perché pensava che l'idea di infinito attuale non potesse essere formulata coerentemente e quindi non potesse trovare cittadinanza nel rigoroso regno della matematica. Ciononostante, secondo quanto lui stesso ne scrive, riuscì a superare il suo «pregiudizio» nei confronti dei numeri transfiniti, perché scoprì che erano indispensabili per un ulteriore sviluppo della matematica. Dati i suoi dubbi iniziali, riuscì a prevedere l'opposizione che gli venne da molte parti, e cercò di prevenirla con argomenti filosofici e teologici, oltre che matematici. Inoltre, quando fu chiamato a rispondere ai suoi critici, seppe propugnare le sue idee con forza notevole. La sua malattia mentale, lungi dallo svolgere un ruolo interamente negativo, può aver contribuito, nelle sue fasi maniacali, all'energia e alla dedizione con cui Cantor si fece promotore della sua teoria.

Georg Ferdinand Ludwig Philipp Cantor nacque il 3 marzo 1845 a Pietroburgo (oggi Leningrado). La madre, Maria Anna Böhm, proveniva da una famiglia di cui facevano parte numerosi valenti musicisti, il più noto dei quali era Joseph Böhm (zio di Maria Anna), direttore di un conservatorio di Vienna e fondatore di una scuola per violinisti che preparò molti fra i maggiori virtuosi dell'epoca. Il padre di Georg, Georg Woldemar Cantor, era un commerciante di successo e un luterano devoto, che inculcò nel figlio profonde convinzioni religiose. Secondo il famoso libro di Eric Temple Bell, *I grandi matematici* (Sansoni, 1966[2]), pubblicato originariamente nel 1937, le insicurezze che Cantor manifestò nella sua vita adulta ebbero origine da un rovinoso conflitto freudiano con il padre, ma le lettere che ci sono rimaste e altri dati di fatto che riguardano il loro rapporto indicano il contrario. Cantor padre sembra essere stato un uomo sensibile, che dedicava particolare attenzione ai figli e manifestava anche un interesse speciale, ma non soffocante, per il benessere e l'educazione del figlio maggiore.

Quando Georg era ancora bambino, la famiglia si trasferì dalla Russia alla Germania, e fu qui che cominciò a studiare matematica. Nel 1868 ricevette il dottorato presso l'Università di Berlino, con una dissertazione sulla teoria dei numeri; due anni più tardi accettò un incarico come *Privatdocent*, ovvero come istruttore, presso l'Università di Halle, una istituzione rispettata ma non prestigiosa, per la matematica, come le Università di Göttingen o di Berlino. Uno dei suoi colleghi ad Halle, Heinrich Eduard Heine, si occupava allora della teoria delle serie trigonometriche, e incoraggiò Cantor a interessarsi del difficile problema dell'unicità di tali serie. Nel 1872, all'età di 27 anni, Cantor pubblicò un saggio, in cui presentava una soluzione molto generale del problema, che conteneva i germi di quella che in seguito sarebbe diventata la teoria degli insiemi transfiniti.

Il problema suggerito da Heine traeva origine dai lavori del matematico francese Jean Baptiste Joseph Fourier. Nel 1822 Fourier aveva dimostrato che il grafico di qualunque curva «ragionevolmente regolare» (una curva, cioè, che presenti al più un numero finito di punti di discontinuità) poteva essere rappresentato in un intervallo come somma di una serie trigonometrica infinita. In altre parole, sovrapponendo un numero infinito di onde sinusoidali e cosinusoidali è possibile approssimare, con qualunque grado di accuratezza desiderato, tutti i punti della curva «ragionevolmente regolare», fatta eccezione per i punti di discontinuità *(si veda l'illustrazione nella pagina seguente)*. Si dice che una simile serie converge alla curva, o alla funzione, fuorché in un numero finito di punti, ovvero «quasi ovunque». Il risultato di Fourier era di grandissima importanza matematica, perché suggeriva la possibilità di rappresentare, o di sostituire, certe funzioni complicate con somme di seni o coseni, molto più semplici da manipolare matematicamente. Al fine di giustificare tale sostituzione, però, era necessaria qualche garanzia del-

Questo ritratto di Cantor e della moglie fu eseguito intorno al 1880, quando Cantor era al culmine della sua carriera. Aveva già cominciato a stupire i matematici dimostrando che l'insieme infinito dei numeri reali, rappresentato dal continuo dei punti di una retta, è più grande dell'insieme infinito di tutte le frazioni. Dimostrò anche la possibilità di definire grandezze infinite, chiamate numeri transfiniti, in grado di descrivere tali differenze. Qualche anno più tardi, Cantor fu colpito da un forte attacco di una psicosi maniaco-depressiva, che finì per porre termine alla sua attività creativa in matematica. La fotografia originale è conservata nella collezione privata di Egbert Schneider.

l'unicità della serie trigonometrica convergente alla funzione. Cantor cominciò a studiare le condizioni nelle quali una serie trigonometrica convergente a una funzione è unica.

Nel 1870 Cantor ottenne il suo primo risultato: se una funzione è continua in un intervallo, la sua rappresentazione in serie trigonometrica è unica. Il suo passo successivo fu quello di rendere meno rigida la condizione della continuità in un intervallo. Supponiamo, per esempio, che la funzione da approssimare mediante serie sia definita come segue: il grafico della funzione è una linea retta parallela all'asse delle x (asse orizzontale), fuorché nel punto $x=1/2$. Per tale valore di x la funzione assume valore 0 anziché 1. Cantor riuscì a dimostrare che, se si rinuncia alla convergenza nel punto $x=1/2$, esiste ancora una sola serie trigonometrica che converge alla funzione data in ogni altro punto. Non esiste altra serie trigonometrica di forma simile che approssimi la funzione. A quel punto fu facile per Cantor estendere il risultato in modo da includere tutte le funzioni con un numero qualunque di discontinuità, che Cantor chiamava punti eccezionali.

Nel 1872 la ricerca di Cantor di un enunciato più generale per il suo teorema di unicità portò alla pubblicazione di una scoperta degna di nota. Scoprì che, se distribuiti in un modo attentamente specificato sull'asse delle x, i punti eccezionali possono essere in numero infinito. Il passo più importante nella dimostrazione del nuovo risultato era la descrizione precisa dell'insieme infinito dei punti eccezionali, e a questo proposito Cantor si rese conto di dover fornire una analisi soddisfacente del continuo dei punti lungo l'asse x. Così lo studio delle serie trigonometriche portò Cantor ad appuntare la propria attenzione sulle relazioni fra punti nel continuo, anziché sui teoremi relativi alle serie trigonometriche.

Cantor considerava assiomatica l'esistenza, per ogni punto di una retta continua, di un numero corrispondente, chiamato un numero reale, al fine di distinguerlo dai numeri «immaginari», che sono multipli di $\sqrt{-1}$. Viceversa, a ogni numero reale corrisponde esattamente un punto su una retta continua. Il problema di descrivere il continuo dei punti su una retta equivale, pertanto, al problema di definire i numeri reali e di esplorarne le proprietà. Uno dei principali risultati del saggio di Cantor del 1872 era una spiegazione rigorosa di tali proprietà.

Gli ostacoli principali per una teoria dei numeri reali sono rappresentati da numeri come π e $\sqrt{2}$, cioè numeri non razionali. (Un numero razionale è un numero esprimibile come quoziente di due interi. Già nell'antichità si era scoperto che $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$ e molte altre radici sono irrazionali.) La legittimità dei razionali non era in dubbio, e pertanto Cantor seguì un'impostazione suggerita da Karl Weierstrass, uno dei suoi maestri all'Università di Berlino. Cantor propose di rappresentare ciascun numero irrazionale mediante una successione infinita di numeri razionali. $\sqrt{2}$, per esempio, può essere rappresentato dalla successione infinita di razionali 1, 1,4, 1,41 e via dicendo: in questo modo tutti gli irrazionali possono essere visti come punti geometrici su una retta numerica, esattamente come i razionali.

Nonostante presentasse indubbi pregi, l'impostazione di Cantor lasciò fortemente perplessi alcuni matematici, in quanto presupponeva l'esistenza di insiemi o successioni di numeri con infiniti elementi. Sin dai tempi di Aristotele, filosofi e matematici hanno rifiutato l'idea di un infinito «completato», soprattutto a causa dei paradossi logici che quella nozione sembra inevitabilmente generare. Galileo, per esempio, sottolineava che, se in matematica fossero ammissibili insiemi infiniti completati, dovrebbero esistere

Una funzione regolare, continua, può essere approssimata con il grado di approssimazione desiderato mediante una serie trigonometrica, cioè una somma di funzioni seno e coseno. Per esempio, una funzione di x che nell'intervallo fra 0 e 1 abbia valore costante 1 (e che quindi in questo intervallo è rappresentata da un segmento di retta parallelo all'asse x, in colore nel disegno) può essere approssimata sovrapponendo più funzioni sinusoidali *(curve in grigio)*: nell'illustrazione sono visualizzati i primi due stadi dell'approssimazione *(curve in nero nei grafici in alto e al centro)*. La serie trigonometrica che approssima la funzione è unica. Anche se il grafico della funzione non è continuo, però, la funzione può spesso essere approssimata da un'unica serie trigonometrica. Per esempio, se la funzione dell'esempio avesse valore costante e uguale a 1 in tutto l'intervallo fra 0 e 1 fuorché nel punto $x = 1/2$, la serie che converge alla curva continua convergerebbe anche alla curva discontinua, fuorché nel punto $x = 1/2$ *(grafico in basso)*. Cantor dimostrò che una funzione può essere approssimata con un'unica serie trigonometrica anche se il numero dei punti di discontinuità è infinito, purché i punti di discontinuità siano distribuiti in modo specifico. Fu pertanto condotto a considerare le proprietà strutturali di insiemi astratti infiniti e gli infiniti modi in cui gli elementi di insiemi infiniti possono venire ordinati.

tanti interi pari quanti sono gli interi pari e dispari insieme. A ciascun intero pari si può far corrispondere in modo unico l'intero che è la sua metà e in questo modo si instaura una corrispondenza biunivoca fra tutti gli elementi di un insieme e tutti gli elementi dell'altro. I teologi come Tommaso d'Aquino (un'altra voce dell'opposizione tradizionale in atto all'infinito completato) consideravano quest'idea come una sfida diretta alla natura unica e assolutamente infinita di Dio.

Per rispondere a queste obiezioni di solito i matematici avevano tracciato una distinzione netta fra l'infinito attuale, considerato come una grandezza completata, e l'infinito visto come potenziale, rappresentato da una somma o da una serie indefinita di numeri, che tende a qualche limite. I matematici erano disponibili a difendere solo l'infinito potenziale. Nel 1831 Carl Friedrich Gauss esprimeva la sua opposizione all'uso degli infiniti attuali in termini che Cantor una volta definì autorevoli. In una lettera a Heinrich Schumacher, Gauss scriveva: «Ma per quanto riguarda la vostra dimostrazione, contesto soprattutto l'uso di una grandezza infinita come *completata*, il che in matematica non è mai ammesso. L'infinito è solo un *façon de parler*, laddove correttamente si deve parlare di limiti.»

Parlando di limiti era possibile evitare i paradossi associati con gli infiniti attuali. Per esempio, aggiungendo altre cifre decimali allo sviluppo decimale di π si può approssimare il valore vero di π con precisione sempre crescente. Gauss insisteva, tuttavia, che non si deve mai ipotizzare di aver fornito tutti i termini dello sviluppo decimale, in modo da determinare esattamente il valore decimale di π. Questo significherebbe «afferrare» un numero infinito di termini nella sua totalità, o in altre parole un insieme attualmente infinito di numeri, una operazione che Gauss rifiutava di ammettere.

Cantor non fu il solo a studiare in modo particolareggiato e rigoroso le proprietà del continuo. Nel 1872, lo stesso anno in cui apparve il saggio di Cantor, anche il matematico tedesco Richard Dedekind pubblicò un'analisi del continuo basata sugli insiemi infiniti. Dedekind articolò un'idea che Cantor, più tardi, rese più precisa: «La retta è infinitamente più ricca di punti-individui di quanto il dominio... dei numeri razionali sia ricco di numeri-individui». Si può vedere questa proprietà in una prospettiva migliore se si considera la distribuzione, su un segmento di retta, dei punti che corrispondono ai numeri razionali, cioè dei punti razionali. Per quanto piccolo, il segmento conterrà sempre un numero infinito di punti razionali. Quindi l'essenza dell'osservazione di Dedekind era che, nonostante la densità dei punti razionali, su un segmento di retta vi è ancora spazio per «impaccare» un numero infinito di punti irrazionali. Punti irrazionali come $\sqrt{2}$ cadono fra i razionali, e quindi l'insieme dei punti razionali, per quanto denso, è ancora pieno di «buchi» e non è continuo.

L'enunciato di Dedekind è coerente con una corretta comprensione del continuo, ma nasconde una grave debolezza. Se qualcuno gli avesse chiesto quanto l'insieme infinito di punti nel continuo fosse più ricco dell'insieme infinito dei numeri razionali, Dedekind non avrebbe risposto. Il contributo principale di Cantor al problema fu pubblicato nel 1874 sul «Journal für die reine und angewandte Mathematik» di August Leopold Crelle, spesso chiamato semplicemente il «Journal di Crelle», forse il più accreditato fra tutti i periodici matematici dell'epoca.

In effetti Cantor prese a prestito il paradosso citato da Galileo e lo trasformò in un mezzo per confrontare le dimensioni di insiemi infiniti. Definì che due insiemi sono equivalenti se è possibile stabilire fra i loro elementi una corrispondenza biunivoca. Immaginate, per esempio, un secchio pieno di biglie nere e colorate, e supponete che qualcuno voglia confrontare il numero delle biglie nere con quello delle biglie colorate. Il modo più semplice è togliere tutte le biglie dal secchio a coppie, una nera e una colorata alla volta. Se è sempre possibile accoppiare una biglia nera e una colorata, i due insiemi sono equivalenti; in caso contrario, le biglie che restano nel secchio costituiscono la base per il confronto.

Applicando lo stesso principio di corrispondenza, Cantor dimostrò che la proprietà considerata paradossale da Galileo era in realtà una proprietà naturale degli insiemi infiniti. L'insieme dei numeri pari è equivalente all'insieme di tutti gli interi positivi, pari e dispari insieme, poiché l'accoppiamento fra gli elementi di ciascun insieme può procedere per sempre, senza omettere alcun elemento dell'uno o dell'altro insieme. Cantor riuscì anche a fornire un metodo ingegnoso e raffinato per mettere in corrispondenza l'insieme di tutti i razionali e l'insieme degli interi *(si veda l'illustrazione in basso nella pagina seguente)*. Cantor definì «numerabile» ogni insieme che potesse essere messo in corrispondenza biunivoca con l'insieme degli interi positivi (il che equivale, in un certo senso, a «contare» gli elementi dell'insieme stesso).

Data la densità dei razionali sulla retta numerica e la relativa dispersione degli interi, può sembrare estremamente controintuitivo che i due insiemi abbiano la stessa dimensione. La mossa successiva di Cantor, tuttavia, fu ancora più sconvolgente: dimostrò infatti che non può esistere

Le dimensioni di due insiemi possono essere confrontate mettendo in corrispondenza gli elementi dell'uno con gli elementi dell'altro. Per esempio, al fine di determinare se in un secchio si trovano più biglie colorate che biglie nere, si possono togliere le biglie a coppie, una colorata e una nera alla volta. Quando non è più possibile estrarre coppie di tal fatta, le eventuali biglie rimaste nel secchio costituiscono la base per il confronto. Cantor applicò il medesimo principio elementare per descrivere in che modo possono essere messe a confronto le dimensioni di insiemi infiniti.

I numeri interi possono essere messi in corrispondenza biunivoca con i numeri pari. Pertanto, benché possa sembrare, intuitivamente, che i numeri interi siano più dei pari, i due insiemi infiniti contengono in effetti lo stesso numero di elementi. Molti altri insiemi infiniti, come l'insieme dei quadrati perfetti moltiplicati per un miliardo, o l'insieme dei cubi, possono essere messi in corrispondenza biunivoca con l'insieme degli interi. Questi tipi di insiemi sono chiamati numerabili.

una simile corrispondenza biunivoca fra l'insieme degli interi e l'insieme di tutti i punti della retta, cioè l'insieme dei reali; in breve, i numeri reali costituiscono un insieme non numerabile. Nel suo saggio del 1874 Cantor diede una dimostrazione un po' complicata di questa affermazione: qui presenterò invece l'idea di fondo di una versione molto più semplice e molto più efficace, che Cantor diede nel 1891.

Cantor inizia la sua dimostrazione supponendo che esista una corrispondenza fra l'insieme dei reali e quello degli interi; poi procede a dimostrare che tale supposizione porta a una contraddizione; ne segue che la supposizione di partenza deve essere falsa e che una tale corrispondenza biunivoca non è possibile. L'argomentazione può essere semplificata esaminando solo i reali compresi fra 0 e 1: se questo insieme di numeri reali è maggiore dell'insieme degli interi, anche l'insieme di tutti i reali sarà maggiore dell'insieme degli interi.

Supponiamo, dunque, che i reali compresi fra 0 e 1 possano essere posti in corrispondenza biunivoca con gli interi. Stabilire una tale corrispondenza equivale a stendere un elenco dei reali, ciascuno dei quali rappresentato come un numero decimale infinito. A questo punto è possibile definire un nuovo numero reale che non è stato incluso nell'elenco. Consideriamo la prima cifra del primo sviluppo decimale nell'elenco: se è 1, scriviamo un 9 nella prima posizione decimale del numero da definire; se non è 1, scriviamo un 1 nella prima posizione decimale del numero da definire. Continuiamo a costruire il nuovo numero modificando la seconda cifra del secondo sviluppo decimale dell'elenco, la terza cifra del terzo sviluppo decimale, e via dicendo. Il numero così costruito deve differire per almeno una cifra decimale da tutti gli altri reali già elencati, ma rappresenterà ancora un numero reale compreso fra 0 e 1. Abbiamo costruito un numero che non compariva nell'elenco dei reali, e pertanto la supposizione che sia possibile elencare tutti i reali porta a una contraddizione.

L'insieme infinito dei numeri razionali (cioè i numeri che possono essere espressi come quoziente di due interi) potrebbe sembrare molto più grande dell'insieme degli interi. Per esempio, tra due interi consecutivi qualunque, come 0 e 1, esistono infiniti numeri razionali. Ciononostante, nel 1874 Cantor dimostrò che i numeri razionali possono essere posti in corrispondenza biunivoca con gli interi. Si possono elencare tutti i possibili razionali in una matrice quadrata come quella indicata nell'illustrazione: ogni razionale può poi essere associato con un intero seguendo il percorso indicato dalle frecce in colore. L'insieme dei razionali risulta dunque numerabile.

Nell'agosto 1874, Cantor sposò Vally Guttmann; i due trascorsero l'estate sulle montagne dell'Harz, dove ebbero modo di incontrare anche Dedekind. Fu un periodo estremamente fruttuoso per le ricerche di Cantor. In precedenza, quello stesso anno, aveva posto a Dedekind, in una lettera, il suo successivo problema: «Sarebbe possibile far corrispondere a una superficie (un quadrato, magari, compresi i suoi lati) una linea retta (magari un intervallo, estremi inclusi) in modo che a ciascun punto della superficie corrisponda un punto della retta, e viceversa?» Cantor pensava che la risposta dovesse essere negativa, ma non sapeva spiegare perché, e anche Dedekind era nella medesima situazione.

Nel 1877, però, Cantor comunicò a Dedekind l'incredibile notizia: contrariamente all'opinione matematica corrente, non è impossibile istituire una corrispondenza biunivoca fra linee e piani. La dimostrazione si basa sulla rappresentazione di ciascun punto di un quadrato mediante una coppia di coordinate decimali. Le rappresentazioni decimali quindi vengono mescolate in un modo ben definito, dando un singolo sviluppo decimale, che viene associato a un punto sul segmento di retta. Il procedimento può essere anche rovesciato *(si veda l'illustrazione a pagina 91)*. Lo stesso Cantor era così impreparato a questo esito, da essere spinto a esclamare «Lo vedo, ma non lo credo!».

Cantor preparò immediatamente un manoscritto, in cui descriveva la sua nuova scoperta, e lo inviò, come il precedente saggio del 1874, al «Journal di Crelle». Nonostante il significato fondamentale del risultato, il manoscritto costituì anche la prima occasione di apertura delle ostilità fra Cantor e il suo maestro, Kronecker. Come direttore del periodico, Kronecker era in condizioni di bloccare la pubblicazione di qualunque articolo, e nel 1877 era così stupito dalla direzione assunta dalle ricerche di Cantor, che prese proprio quella decisione. Benché Cantor avesse inviato il suo manoscritto il 12 luglio, non fu fatto nulla per prepararlo per la pubblicazione, e non apparve nel volume della rivista per l'anno 1877. Cantor, sospettando l'intervento di Kronecker, scrisse una lettera dai toni amari a Dedekind, lamentandosi per il trattamento subito dal suo articolo, ed esprimendo la propria inclinazione a ritirarlo dalla rivista. Dedekind, richiamandosi all'esperienza fatta in prima persona in situazioni del genere, persuase Cantor ad aspettare. Dedekind, in effetti, aveva ragione: l'articolo apparve nel volume del 1878, ma Cantor si era tanto offeso per l'incidente, che si rifiutò di pubblicare ancora qualcosa sul «Journal di Crelle».

La controversia sorta fra Cantor e Kronecker fu oscurata dall'animosità personale, ma alla radice stavano concezioni filosofiche della matematica fortemente divergenti. Quelle concezioni si riflettono ancora nel dibattito fra costruttivisti e formalisti. Kronecker, un precursore del costruttivismo, è ben noto per un detto che condensa bene l'essensa della sua posizione: «Dio ha creato gli interi; tutto il resto è opera dell'uomo». Con questo spirito sosteneva la necessità di costruire tutta la matematica dagli interi e da combinazioni aritmetiche finite degli interi. Agli inizi degli anni settanta dell'Ottocento, cominciò a rifiutare tutti i processi di limite dell'analisi tradizionale, e si oppose a tutti i tentativi di definire oggetti matematici in termini di limiti. Così anche i numeri irrazionali, che erano stati accettati dai matematici per secoli, dovevano essere esclusi dalla matematica, a meno che non si trovasse un metodo per costruirli (come erano stati costruiti i razionali) a partire dagli interi.

Cantor, che aveva scritto due saggi importanti sotto la guida di Kronecker mentre era studente all'Università di Berlino, conosceva bene la posizione estremistica del maestro e non era insensibile ai suoi pregi. Garantiva infatti alle dimostrazioni matematiche la massima sicurezza e correttezza e poteva essere applicata come un elemento correttivo alle speculazioni matematiche più sfrenate. Ciononostante, Cantor sosteneva che accettare la posizione di Kronecker significava mettere fuori legge molti fra gli sviluppi più promettenti della matematica; inoltre, poteva anche mettere a carico delle ricerche innovative scrupoli metodologici un po' miopi e alla fin fine sterili.

La definizione dei numeri irrazionali data da Cantor nel saggio del 1872 equivaleva ad accettare l'esistenza di insiemi infiniti attuali. Cantor abbracciò una posizione formalistica sull'esistenza degli irrazionali e sostenne che l'unico fondamento su cui poteva essere giudicata la loro legittimità in matematica era la loro coerenza formale e interna. «Nell'introdurre nuovi numeri - scrisse una volta - la matematica è obbligata solamente a darne definizioni, grazie alle quali... possano essere chiaramente distinti l'uno dall'altro. Non appena un numero soddisfa tutte queste condizioni, deve essere considerato come esistente e reale in matematica.»

Questo punto di vista sugli irrazionali fu cruciale per la giustificazione che Canton diede all'introduzione dei numeri transfiniti. Nel saggio del 1872 aveva definito insiemi di punti eccezionali introducendo il concetto di punto limite. Il numero irrazionale $\sqrt{2}$, per esempio, è un punto limite della successione 1, 1,4, 1,41 ecc. Più in generale, un punto è un punto limite di un insieme infinito se esiste un numero infinito di membri dell'insieme che giacciono a distanza arbitrariamente piccola da quel punto.

Dato un insieme infinito P, Cantor definiva l'insieme derivato P^1 come l'insieme di tutti i punti limite di P. Analogamente, se anche P^1 era un insieme infinito, il suo insieme derivato P^2 era definito come l'insieme di tutti i punti limite di P^1. Cantor dimostrò che la relazione di sottoinsieme determina un ordinamento naturale per gli insiemi: ogni elemento di P^2 è elemento anche di P^1, e pertanto P^2 è un sottoinsieme di P^1; analogamente, P^3 è un sottoinsieme di P^2, e via di seguito.

Può darsi che, per qualche intero finito n, l'insieme derivato P^n sia un insieme finito; se questa condizione sussiste, l'insieme infinito di punti P che dà luogo all'insieme P^n è esattamente l'insieme dei punti eccezionali necessario per dimostrare la versione generalizzata del teorema di Cantor sull'unicità della rappresentazione di funzioni mediante serie trigonometriche. Può anche accadere, tuttavia, che nessuno degli insiemi derivati nella successione $P^1, P^2, P^3, \ldots$ sia finito. Cantor sostenne che in questo caso ha ancora senso considerare l'insieme dei punti comuni a tutti gli insiemi derivati $P^1, P^2, P^3, \ldots, P^n, \ldots$ Indicò con P^∞ l'insieme dei punti comuni a tutti questi insiemi derivati, e nel 1880 cominciò a parlare dell'indice soprascritto ∞ in termini di simbolo transfinito. Inoltre, se P^∞ fosse un insieme infinito di punti, si potrebbe formare il suo insieme derivato $P^{\infty+1}$, che a sua volta potrebbe dare il via a una intera successione di insiemi derivati, $P^{\infty+2}$ e via dicendo.

Cantor avrebbe potuto aggiungere che gli indici soprascritti ∞, $\infty+1$, $\infty+2$ ecc. costituiscono effettivamente un nuovo tipo di numeri, ma a tutta prima non lo fece. Nel 1872 era stato attento a discutere gli irrazionali solo in termini di successioni di razionali e, analogamente, agli

L'insieme dei numeri reali, rappresentato dal continuo dei punti su una retta, non è numerabile. Se lo fosse, i reali compresi, per esempio, fra 0 e 1 potrebbero essere posti in corrispondenza biunivoca con gli interi. Ciascuno dei reali nell'elenco può essere rappresentato da uno sviluppo decimale infinito. (I decimali infiniti come 0,5000... debbono essere rappresentati da un decimale infinito equivalente, come 0,4999...) Non importa quanti siano i decimali infiniti elencati; si può sempre costruire un nuovo decimale che definisce un numero reale non compreso nell'elenco: si scrive un 9 nella prima posizione decimale del numero costruito, se la prima cifra decimale del primo reale dell'elenco è 1; altrimenti la prima cifra decimale del numero costruito sarà un 1. Si procede modificando nello stesso modo la seconda cifra decimale del secondo reale nell'elenco, la terza cifra del terzo reale, e via di seguito. Lo sviluppo decimale così costruito rappresenta un numero reale compreso fra 0 e 1, ma che differisce per almeno una posizione decimale da tutti i reali compresi nell'elenco. La supposizione che i numeri reali possano essere posti in corrispondenza biunivoca con i numeri naturali porta a una contraddizione e quindi va abbandonata. L'idea di fondo impiegata in questa dimostrazione va sotto il nome di «metodo di diagonalizzazione».

inizi considerò i simboli ∞, $\infty+1$, $\infty+2$ ecc. solo come etichette per identificare insiemi. Nel 1883, però, abbandonò la sua reticenza e li presentò come numeri transfiniti, una estensione autonoma e sistematica dei numeri naturali.

Il motivo immediato per la loro introduzione, sosteneva Cantor, era legato al fatto che erano necessari per un progresso ulteriore della teoria degli insiemi e dello studio dei numeri reali. Ciononostante, per rispondere a critici come Kronecker, Cantor introdusse nella discussione il suo punto di vista filosofico formalistico: una volta riconosciuta la coerenza interna dei numeri transfiniti, non si poteva rifiutar loro un posto accanto a tutti gli altri membri accettati (ma un tempo assai discussi) della famiglia matematica, come gli irrazionali. Formulando una teoria dell'infinito che riusciva a evitare i ben noti paradossi matematici, Cantor credeva di aver rimosso l'unica obiezione razionale che i matematici potessero avanzare, per rifiutarsi di considerare nel loro lavoro l'infinito attuale.

I numeri transfiniti alla fine introdotti da Cantor sono noti nella notazione adot-

tata dallo stesso Cantor nei suoi ultimi anni, basata sulla prima lettera dell'alfabeto ebraico ℵ (aleph). Gli aleph indicano la cardinalità, ovvero il numero degli elementi, di un insieme infinito e quindi le equivalenze fra insiemi infiniti, dimostrate da Cantor negli anni settanta del secolo scorso, vengono formulate spesso in termini dei cardinali transfiniti, gli aleph. Ha un notevole interesse storico pertanto il fatto che i primi numeri transfiniti introdotti non fossero cardinali, ma ordinali.

Un numero ordinale è definito dal suo ordine, ovvero dalla sua posizione all'interno di un elenco. Il numero ordinale associato con un insieme finito corrisponde al cardinale dell'insieme. Per esempio, qualunque insieme che abbia cinque elementi (cioè qualunque insieme il cui cardinale sia cinque) può essere pensato, in un certo senso, come il successore immediato di qualunque insieme che abbia quattro elementi. In altre parole, anche l'ordinale dell'insieme è cinque: è il quinto insieme in un elenco di insiemi. L'ordinale di un insieme infinito, tuttavia, deve essere ben distinto dal suo cardinale. Cantor dimostrò che era possibile costruire un numero infinito di insiemi infiniti che possiedono ordinali diversi, ma lo stesso cardinale. In effetti, Cantor in seguito scoprì che è possibile trasformare questa proprietà degli insiemi infiniti in un criterio per distinguerli dagli insiemi finiti. Un insieme è finito solamente se il suo cardinale e il suo ordinale coincidono.

Cantor sottolineò che l'ordinale di una successione di insiemi finiti di dimensione crescente 1, 2, 3 ecc. si basa sulla somma ripetuta di unità. Non esiste un massimo ordinale associato con la sequenza degli insiemi finiti, ma come è possibile definire π come il limite di una successione di numeri razionali, senza con ciò supporre che π stesso sia un numero razionale, così, pensava Cantor, è corretto definire un nuovo numero ordinale transfinito ω come il primo numero che segue tutta la successione dei comuni ordinali 1, 2, 3 ecc. Una volta definito ω è possibile generare, mediante ripetuta somma di unità, altri ordinali transfiniti, $\omega+1, \omega+2, \omega+3$ ecc. Poiché nemmeno questa successione ha un elemento massimo, si può immaginare un altro ordinale $\omega+\omega$, ovvero 2ω, primo ordinale che segue tutta la successione $\omega+1$, $\omega+2$, $\omega+3$ ecc. Applicando ripetutamente questi due principi di generazione, Cantor riuscì a definire una gerarchia di ordinali transfiniti progressivamente sempre più grandi *(si veda l'illustrazione a pag. 92)*.

Come si può distinguere l'ordinale ω, poniamo, dall'ordinale $\omega+1$? La differenza è determinata dall'ordine degli elementi all'interno degli insiemi a cui ω e $\omega+1$ corrispondono. Per esempio, l'insieme dei numeri naturali presi nel loro ordine familiare (1, 2, 3,...) ha l'ordinale ω, che rappresenta tutta la successione dei naturali nel loro ordine usuale. L'insieme dei naturali presi nella successione riordinata (2, 3, 4, ..., 1), invece, o l'insieme di tutti i naturali nella successione (10, 30, 40, ..., 20) ha l'ordinale $\omega+1$. In altre parole, la distinzione dipende dall'ordine consecutivo degli elementi della successione e dalla posizione della lacuna infinitamente lunga rappresentata dai puntini ...; se un numero viene spostato alla fine della successione, l'ordinale della nuova successione è $\omega+1$. La successione (2, 4, 6, ..., 1, 3, 5, ...) ha due lacune infinite, quindi il suo ordinale sarà $\omega+\omega$, cioè 2ω. Si noti che tutti questi insiemi hanno lo stesso numero di elementi, cioè i loro elementi possono essere posti in corrispondenza biunivoca fra di loro e con i numeri interi positivi. Pertanto il loro cardinale deve essere sempre il medesimo, anche se i loro ordinali sono molto diversi.

La probabilità di scegliere a caso, dal continuo dei numeri reali, un punto che rappresenta un numero razionale illustra il confronto di dimensioni fra i due insiemi numerici. La probabilità è il rapporto fra il numero dei punti razionali e il numero totale dei punti in un certo intervallo. Qui l'intervallo compreso fra 0 e 1 è rappresentato dalla circonferenza di una ruota della fortuna. (Sulla ruota 0 e 1 coincidono.) Si assume che facendo girare la ruota e lasciandola quindi fermare si possa scegliere casualmente qualunque punto della circonferenza. I punti che rappresentano i razionali sono infinitamente densi, nel senso che, in un arco qualunque compreso fra due razionali deve esistere un numero infinito di punti razionali. Sono stati indicati su tale ruota alcuni di tali punti. Ciononostante, l'insieme di tutti i punti della circonferenza è infinitamente più grande dell'insieme dei punti razionali: la probabilità che la ruota si fermi su un punto razionale è nulla. Più precisamente, la probabilità è più piccola di qualunque quantità positiva specificabile a priori.

Una volta definiti gli ordinali transfiniti, Cantor procedette a descriverne le proprietà aritmetiche. Una differenza importante fra i numeri transfiniti e i numeri ordinari riguarda la proprietà commutativa dell'addizione e della moltiplicazione. Per due numeri finiti qualunque A e B la proprietà commutativa dice che $A+B$ è uguale a $B+A$ e $A \times B$ è uguale a $B \times A$. Quando si definiscono l'addizione e la moltiplicazione in modo da includere i numeri transfiniti, però, la proprietà commutativa, in generale, non è più valida. Per esempio, $\omega+2$, che rappresenta la successione (1, 2, 3, ..., 1, 2) non è uguale a $2+\omega$, che rappresenta la successione (1, 2, 1, 2, 3, ...).

Benché nel caso degli insiemi finiti numero cardinale e ordinale dell'insieme coincidano, la distinzione tra i due tipi di numeri servì a spiegare perché l'applicazione del concetto di numero a un insieme infinito potesse provocare confusione e generare paradossi. Poiché i concetti di numero ordinale e di numero cardinale sono fondamentalmente distinti, nel caso degli insiemi infiniti, qualunque argomentazione che discuta il numero associato con un insieme infinito senza tracciare quella distinzione è inevitabilmente condannato all'ambiguità. Pertanto non è legittimo estendere le proprietà, apparentemente prive di ambiguità, degli insiemi finiti agli insiemi infiniti, come invece Galileo e altri avevano voluto fare.

Nonostante i risultati raggiunti da Cantor negli anni ottanta del secolo scorso, rimaneva ancora aperto un serio interrogativo. Il problema della cardinalità (o della «potenza», secondo la terminologia iniziale di Cantor) del continuo dei numeri reali attendeva ancora una risposta. Si ricordi che nell'articolo del 1883 Cantor aveva definito la successione degli ordinali transfiniti in base a due principi di generazione. Al fine di introdurre divisioni naturali nella successione, aveva aggiunto un terzo principio. Consideriamo l'insieme di tutti gli interi finiti, che Cantor chiamava la prima classe numerica. La sua potenza (la sua cardinalità) è maggiore della potenza di qualunque elemento dell'insieme. Analogamente, notò, si potrebbe considerare l'insieme di tutti gli ordinali transfiniti corrispondenti a insiemi infiniti numerabili, ovvero in altre parole gli insiemi la cui potenza è identica alla potenza dell'insieme di tutti gli interi. Cantor chiamò seconda classe numerica questo insieme di ordinali transfiniti. Si ha che la potenza della seconda classe numerica è maggiore della potenza di qualunque ordinale transfinito membro dell'insieme. In breve, la seconda classe numerica è un insieme non numerabile. Benché non riuscisse mai a dimostrarlo, Cantor era convinto che la potenza della seconda classe numerica fosse equivalente alla potenza del continuo dei reali.

Questa congettura è divenuta famosa con il nome di «ipotesi del continuo di Cantor» e non è mai stata dimostrata. Nel 1963 Paul J. Cohen, della Stanford University, a partire dai lavori di Kurt Gödel dell'Institute for Advanced Study di Princeton, ha dimostrato che, benché sia coerente con gli assiomi di una versione standard della teoria degli insiemi, l'ipotesi del continuo è anche indipendente da essi. In effetti, l'ipotesi del continuo svolge per la teoria degli insiemi lo stesso ruolo che il postulato delle parallele di Euclide ha rivestito per la geometria. Assumendo vera o falsa l'ipotesi del continuo si possono costruire versioni diverse della teoria degli insiemi, esattamente come assumendo vero o falso il postulato delle parallele di possono costruire la geometria euclidea o le geometrie non euclidee (si veda l'articolo *La teoria non cantoriana degli insiemi* di Paul J. Cohen e Reuben Hersh in «Le Scienze», n. 1, settembre 1968).

I punti nel piano possono essere messi in corrispondenza biunivoca con i punti di una retta. Ciascun punto del piano è rappresentato da una coppia di decimali infiniti, suddivisi in gruppi. Ogni cifra diversa da 0 nello sviluppo decimale inizia un nuovo gruppo. I gruppi vengono combinati disponendoli alternativamente in un unico decimale infinito, che rappresenta un punto sulla retta. Il procedimento può essere invertito. Una dimostrazione analoga dice che il numero dei punti in qualsiasi spazio dimensionale finito è equivalente al numero dei punti di una retta.

L'insuccesso dei suoi sforzi per dimostrare l'ipotesi del continuo provocò in Cantor tensione e ansia. Agli inizi del 1884 pensò di aver trovato una dimostrazione, ma pochi giorni più tardi si ricredette completamente e pensò di poter dimostrare la falsità dell'ipotesi. Alla fine si rese conto di non compiere alcun progresso. Per tutto quel periodo dovette resistere all'opposizione crescente e alle minacce di Kronecker, che - a quanto andava dicendo - stava preparando un articolo che avrebbe dimostrato come «i risultati della teoria moderna delle funzioni e della teoria degli insiemi non avessero alcuna effettiva importanza».

Subito dopo, nel maggio 1884, Cantor accusò il primo dei suoi seri esaurimenti nervosi. La frustrazione per la mancanza di progressi negli studi sull'ipotesi del continuo o la tensione per gli attacchi di Kronecker possono aver favorito l'innescarsi della crisi, ma oggi sembra chiaro che quegli eventi avevano poco a che vedere con la causa sottostante. Il malessere si evidenziò con una velocità impressionante e durò più di un mese. A quell'epoca si riconosceva come un sintomo solo la fase maniacale della psicosi maniaco-depressiva; quando Cantor si «riprese», alla fine del giugno 1884, ed entrò nella fase depressiva, si lamentava che gli mancavano l'energia e l'interesse per ritornare al rigoroso pensiero matematico. Si accontentava di occuparsi di questioni amministrative di scarso rilievo all'università e non si sentiva quasi capace d'altro.

Anche se alla fine ritornò alla matematica, Cantor cominciò a essere sempre più assorbito da altri interessi. Intraprese uno studio della storia e della letteratura inglese, e si lasciò coinvolgere in una discussione accademica, vista con molta serietà da molti, all'epoca: l'ipotesi che Francesco Bacone fosse l'autore delle commedie di Shakespeare. Cantor tentò anche, senza successo, di insegnare filosofia anziché matematica, ed entrò in corrispondenza con numerosi teologi che erano rimasti interessati dalle conseguenze filosofiche delle sue teorie sull'infinito. La corrispondenza ebbe un particolare significato per Cantor, poiché si venne convincendo che i numeri transfiniti gli fossero stati mandati come un messaggio da Dio. Era ansioso che le sue idee venissero studiate con attenzione dai teologi, perché il suo concetto matematico dell'infinito potesse essere riconciliato con la dottrina della Chiesa.

Cosa ancor più importante, Cantor si prodigò per creare una società professionale per la promozione della matematica in Germania, la Deutsche Mathematiker-Vereinigung. Era convinto che la sua carriera fosse stata fortemente danneggiata dal rifiuto prematuro e pregiudiziale del

$$1 \rightarrow \{1\}$$
$$2 \rightarrow \{1, 2\}$$
$$3 \rightarrow \{1, 2, 3\}$$
$$\vdots$$
$$\omega \rightarrow \{1, 2, 3, \ldots\}$$
$$\updownarrow\ \updownarrow\ \updownarrow$$
$$\{2, 4, 6, \ldots\}$$
$$\omega + 1 \rightarrow \{2, 3, 4, \ldots, 1\}$$
$$\updownarrow\ \updownarrow\ \updownarrow\ \ \ \updownarrow$$
$$\{1, 3, 4, \ldots, 2\}$$
$$\omega + 2 \rightarrow \{3, 4, 5, \ldots, 1, 2\}$$
$$\updownarrow\ \updownarrow\ \updownarrow\ \ \ \updownarrow\ \updownarrow$$
$$\{1, 4, 5, \ldots, 2, 3\}$$
$$\vdots$$
$$\omega + \omega = 2\omega \rightarrow \{1, 3, 5, 7, \ldots, 2, 4, 6, 8, \ldots\}$$
$$\updownarrow\ \updownarrow\ \updownarrow\ \updownarrow\ \ \ \updownarrow\ \updownarrow\ \updownarrow\ \updownarrow$$
$$\{1, 2, 5, 6, \ldots, 3, 4, 7, 8, \ldots\}$$

$$2\omega + 1$$
$$2\omega + 2$$
$$\vdots$$
$$2\omega + \omega = 3\omega$$
$$\vdots$$
$$\omega \times \omega = \omega^2$$
$$\omega^2 + 1$$
$$\vdots$$
$$\omega^2 + \omega$$
$$\omega^2 + \omega + 1$$
$$\vdots$$
$$\omega^2 + \omega + \omega = \omega^2 + 2\omega$$
$$\vdots$$
$$\omega^2 + \omega \times \omega = 2\omega^2$$
$$2\omega^2 + 1$$
$$\vdots$$
$$\omega \times \omega^2 = \omega^3$$
$$\omega^3 + 1$$
$$\vdots$$
$$a\omega^n + b\omega^{n-1} + \ldots + z$$
$$\vdots$$
$$\omega^\omega$$
$$\omega^\omega + 1$$
$$\vdots$$
$$\omega^\omega + \omega$$
$$\vdots$$
$$\omega^\omega + \omega^\omega = 2\omega^\omega$$
$$\vdots$$
$$\omega \times \omega^\omega = \omega^{\omega+1}$$
$$\vdots$$
$$\omega^{\omega+2}$$
$$\vdots$$
$$\omega^{\omega+\omega} = \omega^{2\omega}$$
$$\vdots$$
$$\omega^{\omega\times\omega} = \omega^{\omega^2}$$
$$\vdots$$
$$\omega^{\omega^\omega}$$
$$\vdots \ \cdot^{\cdot^{\cdot}}$$
$$\omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}}$$
$$\vdots$$

I numeri ordinali transfiniti sono identificati dal loro ordine, ovvero dalla loro posizione, in un elenco. L'elenco viene generato in base a due principi. In primo luogo, ciascun nuovo ordinale transfinito è derivato da quello che lo precede immediatamente per addizione di una unità, esattamente come se si «contasse» oltre l'ordinale transfinito ω che è associato con l'insieme dei numeri interi disposti nel loro ordine naturale. In secondo luogo, ogni qualvolta esiste una successione di ordinali transfiniti priva di un massimo, si definisce un nuovo ordinale transfinito come numero immediatamente più grande di tutti gli altri. I nuovi numeri vengono elencati subito dopo una serie verticale di puntini; per esempio, 2ω è l'ordinale transfinito immediatamente più grande di tutti i numeri $\omega+1$, $\omega+2$, e via dicendo. Gli ordinali transfiniti si distinguono per la posizione delle lacune *(gruppi orizzontali di puntini)* all'interno degli insiemi a cui corrispondono gli ordinali. Così nel diagramma sono dati due esempi di insiemi che corrispondono ai numeri ordinali ω, $\omega+1$, $\omega+2$ e 2ω. Tutti gli insiemi infiniti rappresentati da un ordinale nell'elenco, tuttavia, hanno lo stesso cardinale, e cioè $\aleph_0$; in altre parole, tutti contengono lo stesso numero di elementi.

suo lavoro da parte dell'«*establishment*» matematico, e sperava che l'organizzazione indipendente potesse incoraggiare i matematici più giovani e potesse costituire un uditorio corretto per idee nuove e magari radicali.

C'era un ultimo elemento della teoria degli insiemi transfiniti che Cantor doveva affrontare, e cioè la natura e lo statuto dei numeri cardinali transfiniti. L'evoluzione delle sue idee in proposito è curiosa, perché i cardinali transfiniti furono l'ultima parte della sua teoria a ricevere sia una definizione rigorosa, sia una nozione speciale. In effetti, è difficile, nella bella luce del senno di poi, ricostruire l'oscurità in cui Cantor deve essersi dibattuto, e fino a questo punto ho discusso i suoi lavori come se Cantor si fosse già reso conto che la potenza di un insieme infinito poteva essere pensata come un numero cardinale. Di fatto, benché sapesse già che è la potenza di un insieme ciò che ne stabilisce l'equivalenza o meno con ogni altro insieme, Cantor inizialmente evitò tutto quello che poteva suggerire la possibilità di interpretare la potenza di un insieme infinito come un numero.

Cominciò ad avvicinare i due concetti nel settembre 1883; tuttavia, non aveva ancora un simbolo per distinguere un cardinale transfinito da un altro. Poiché aveva già adottato il simbolo ω per indicare il più piccolo ordinale transfinito, è chiaro che per il primo sviluppo concettuale della teoria cantoriana degli insiemi gli ordinali debbono essere stati molto più importanti dei cardinali. Quando infine Cantor introdusse un simbolo per il primo cardinale transfinito, prese a prestito la notazione da quella già usata per gli ordinali transfiniti: il primo cardinale transfinito veniva indicato con $\bar{\omega}$.

Cantor non ricorse agli aleph fino al 1893. Allora il matematico italiano Giulio Vivanti stava preparando una presentazione generale della teoria degli insiemi, e Cantor si rese conto che era giunto il momento di adottare una notazione standard. Scelse gli aleph per i cardinali transfiniti, perché pensava che l'alfabeto greco e latino fossero già fin troppo usati in matematica per altri fini. I suoi nuovi numeri meritavano un simbolo distinto e univoco. Così scelse la lettera $\aleph$, che era facilmente disponibile fra i caratteri dei compositori tedeschi. La scelta fu particolarmente indovinata, come Cantor stesso si compiaceva di ammettere, perché la prima lettera dell'alfabeto ebraico era anche un simbolo per il numero 1. Poiché i cardinali transfiniti erano a loro volta unità infinite, l'aleph poteva essere preso a rappresentare un nuovo inizio per la matematica. Cantor indicò con $\aleph_0$ (aleph zero) il cardinale della prima classe numerica, quello che in precedenza aveva indicato con $\bar{\omega}$; il cardinale della seconda classe era indicato invece con $\aleph_1$ (aleph uno).

Cantor portò i suoi ultimi contributi importanti alla teoria degli insiemi in un paio di articoli pubblicati nel 1895 e nel 1897. In un saggio presentato poco tempo prima del primo convegno della Deutsche Mathematiker-Vereinigung nel 1891 aveva dimostrato che il cardinale di qualunque insieme è sempre più piccolo del cardinale dell'insieme di tutti i suoi sottoinsiemi. (Una versione della dimostrazione è data nell'illustrazione della pagina a fronte.) Qualche anno più tardi ne trasse un corollario: il cardinale del continuo è uguale a un cardinale che Cantor indicava come $2^{\aleph_0}$. Sperava che questo risultato portasse presto a una soluzione dell'ipotesi del continuo, perché a quel punto l'ipotesi poteva essere formulata in termini algebrici molto chiari: $2^{\aleph_0} = \aleph_1$.

Le argomentazioni nella dimostrazione cantoriana relativa al cardinale dell'insieme dei sottoinsiemi, tuttavia, portarono a conclusioni di gran lunga differenti. La più importante fu ottenuta da Bertrand Russell nel 1903: Russell dimostrò che nella teoria degli insiemi poteva essere derivato un paradosso, considerando tutti gli insiemi che non sono membri di se stessi. Il paradosso di Russell faceva pensare che nella definizione cantoriana di insieme vi fosse qualche cosa di fondamentalmente errato e le conseguenze del paradosso di Russell hanno costituito un problema importante per la logica matematica del ventesimo secolo. Ciononostante, questi sviluppi non hanno sovvertito alcuno dei risultati principali ottenuti da Cantor nell'aritmetica transfinita.

Sfortunatamente nel 1903 Cantor soffriva sempre più frequentemente di attacchi di depressione maniacale, e non si è trovato alcun documento da cui si possa desumere che Cantor venisse a conoscenza del risultato di Russell. In effetti, il suo stato di salute spinse Cantor a richiedere un permesso all'Università di Halle per l'autunno del 1899, permesso che gli fu accordato. Nel novembre dello stesso anno notificò al Ministero tedesco per la cultura la sua intenzione di lasciare completamente l'insegnamento. Si sarebbe accontentato di una posizione modesta in una biblioteca, purché il suo stipendio non venisse diminuito. Tra le sue qualifiche, sottolineava le pubblicazioni sulla questione shakespeariana e concludeva con la richiesta straordinaria di una risposta da parte del ministero nell'arco di due giorni. Se non gli fosse stata offerta alcuna alternativa all'insegnamento, scriveva, allora, in quanto nato in Russia, avrebbe cercato di entrare nel corpo diplomatico russo.

Sembra che la richiesta non abbia avuto seguito, ma Cantor non entrò al servizio dell'imperatore Nicola II. Ciononostante, l'episodio corrisponde perfettamente all'andamento generale del suo comportamento nel 1884, quando aveva preso in seria considerazione l'idea di abbandonare la matematica per la filosofia, dopo il primo forte esaurimento. Fu ricoverato ancora per depressione maniacale verso la fine del 1899, nei trimestri invernali del 1902 e del 1903 e quindi altre volte per periodi sempre più lunghi e frequenti. Morì per un attacco cardiaco il 6 gennaio 1918, nella Halle Nervenklinik.

Esiste un nesso importante fra la malattia mentale di Cantor e la sua matematica. Taluni documenti fanno pensare che a volte la malattia gli lasciasse periodicamente respiro dai problemi quotidiani, consentendogli di seguire le sue idee matematiche nella solitudine dell'ospedale o nella quiete domestica. La malattia forse sostenne anche la sua convinzione che i numeri transfiniti gli fossero stati comunicati da Dio. Dopo un lungo periodo di ricovero nel 1908, ebbe a scrivere a un amico a Göttingen, il matematico inglese Grace Chisholm Young. La sua depressione maniacale, come egli stesso la descriveva, aveva preso un corso particolarmente fecondo: «Un destino peculiare, che grazie a Dio non mi ha affatto distrutto, ma in realtà mi ha reso più forte intimamente, più felice e più animato da una gioiosa attesa di quanto non mi sia mai accaduto negli ultimi due anni, mi ha tenuto a lungo lontano da casa - potrei anche dire lontano dal mondo... Nel mio lungo isolamento né la matematica né, in particolare, la teoria dei numeri transfiniti si sono mai assopite o sono rimaste in ozio dentro di me.»

Altrove Cantor descriveva la sua convinzione della verità della sua teoria in termini quasi religiosi: «La mia teoria si regge salda come una roccia; ogni freccia diretta contro di essa ritornerà rapidamente a chi l'ha lanciata. Come lo so? Perché l'ho studiata sotto tutti gli aspetti per molti anni, perché ho esaminato tutte le obiezioni che siano mai state mosse contro i numeri infiniti e soprattutto perché l'ho seguita fino alle sue radici, per così dire, fino alla prima causa infallibile di tutto il creato.»

Le generazioni successive potranno mettere da parte la filosofia, guardare con diffidenza i frequenti riferimenti a San Tommaso e ai padri della Chiesa, passare sopra alle enunciazioni metafisiche e tralasciare completamente le radici religiose della tarda fede di Cantor nella verità assoluta della sua teoria. Ma tutti questi fattori hanno contribuito alla sua risoluzione di non abbandonare i numeri transfiniti. L'opposizione sembra aver rafforzato la sua determinazione. La sua sopportazione, non meno di tutti gli altri suoi contributi scientifici, ha fatto sì che la teoria degli insiemi sopravvivesse ai primi anni di dubbio e di rigetto per fiorire infine come una forza vigorosa e rivoluzionaria all'interno del pensiero scientifico del ventesimo secolo.

Si può costruire una successione infinita di insiemi, ciascuno più grande del precedente, considerando l'insieme di tutti i sottoinsiemi di un insieme dato. Si può utilizzare un'ingegnosa variante del metodo di diagonalizzazione di Cantor per mostrare come, assumendo una corrispondenza biunivoca *f* fra un insieme qualunque *M* e l'insieme di tutti i suoi sottoinsiemi *N*, sia possibile costruire un sottoinsieme *S* non incluso nella corrispondenza, qualunque sia *f*. Consideriamo l'insieme finito *M* che comprende un disco rosso, un disco azzurro e un disco verde. L'insieme ha otto sottoinsiemi (compreso l'insieme vuoto Φ, che non contiene alcun elemento). Definiamo come *S* l'insieme di tutti gli elementi *m* di *M* che non sono elementi del sottoinsieme *f(m)* a cui corrisponde *m*. Nell'esempio illustrato, *S* comprende solo il disco azzurro. Poiché *S* è un sottoinsieme di *M* e poiché si assume che *f* sia una corrispondenza biunivoca, deve esistere un elemento *a* di *M* a cui corrisponde *S*, o in altre parole un elemento *a* tale che *f(a)* sia identico a *S*. O *a* è un elemento di *S* o non lo è. Se è un elemento di *S*, deve essere elemento anche di *f(a)*, poiché *f(a)* è uguale a *S*; d'altra parte, se *a* è elemento di *S*, non può essere elemento di *f(a)* per la definizione di *S*. Pertanto *a* non può essere elemento di *S*. Di nuovo, però, per la definizione di *S*, se *a* non è elemento di *S*, deve essere elemento di *f(a)* e, poiché *f(a)* è uguale a *S*, *a* deve essere elemento anche di *S*. Non importa dunque quale sia lo stato di *a*: la supposizione che *M* possa essere messo in corrispondenza biunivoca con l'insieme di tutti i suoi sottoinsiemi porta a una contraddizione e va abbandonata. Nello stesso modo si dimostra che, anche se un insieme è infinito, l'insieme di tutti i suoi sottoinsiemi è più grande dell'insieme di partenza. Si può dunque costruire una successione di insiemi infiniti sempre più grandi, formando l'insieme *N* di tutti i sottoinsiemi di un insieme infinito *M*, poi formando l'insieme *P* di tutti i sottoinsiemi di *N*, ecc. La successione non ammette un massimo.

I nastri presinaptici nei recettori visivi

In posizione strategica vicino alle sinapsi che i fotorecettori stabiliscono con le cellule bipolari e orizzontali, ne modulano i messaggi, controllando la liberazione delle vescicole sinaptiche

di Ruggero Pierantoni

La luce, dopo esser passata attraverso il cristallino e aver percorso il corpo vitreo, incontra la retina. La retina è un plesso nervoso di tipo laminare disteso sulla coroide, la quale è a sua volta a contatto con la superficie interna della sclera che è il contenitore semirigido di tutta l'attrezzatura ottico-nervosa dell'occhio propriamente detto. Essa è quindi una sottile trama nervosa il cui spessore raggiunge circa un terzo di millimetro. La sua faccia in contatto diretto con il corpo vitreo è percorsa dai grandi vasi che la irrorano e che costituiscono gli elementi di riconoscimento essenziali nello studio oftalmologico del fondo oculare. L'altra faccia, quella in contatto quasi diretto con la sclera, è coperta da una popolazione fitta e molto omogenea di cellule specializzate: i fotorecettori.

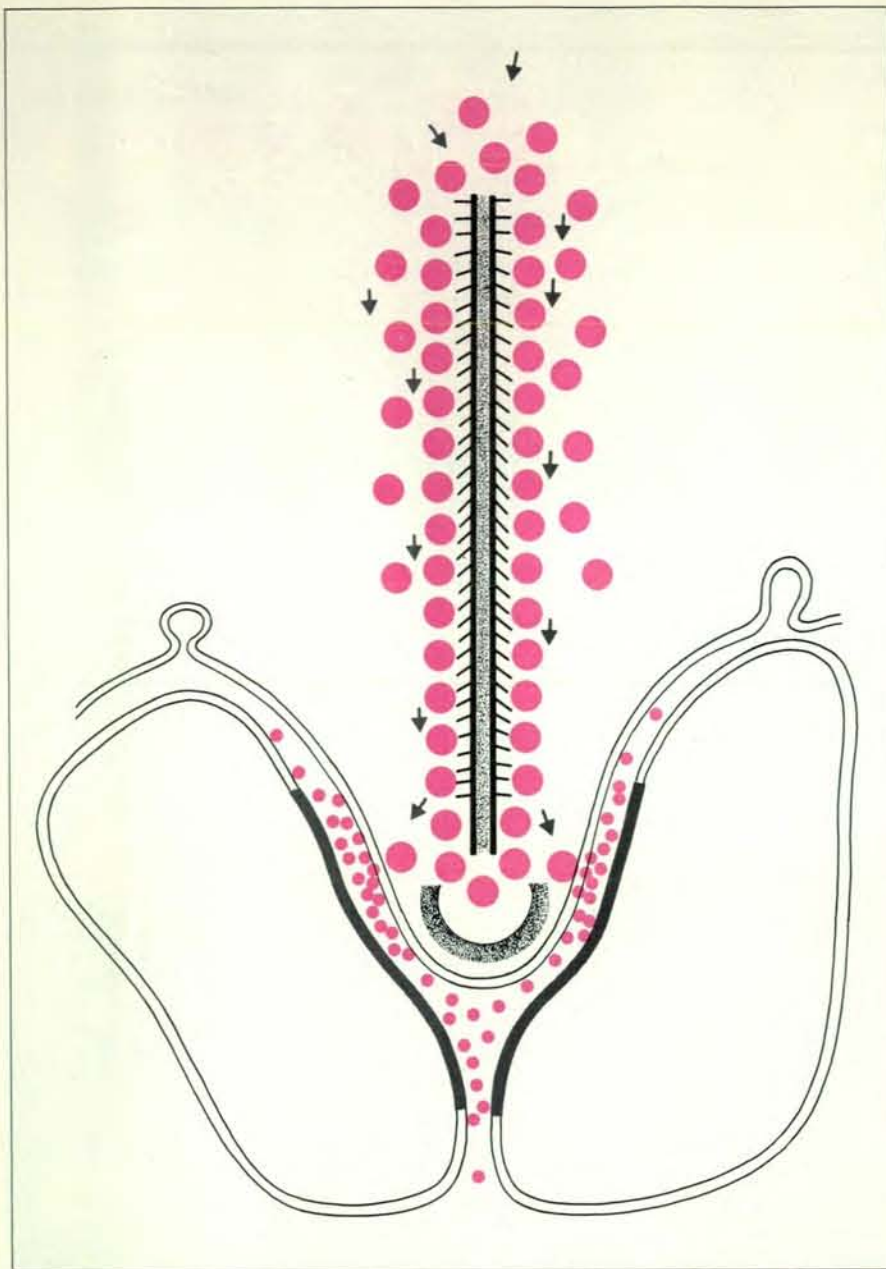
In questa terminazione di cono di retina di tartaruga, fotografata al microscopio elettronico dall'autore presso il Department of Biological Sciences del California Institute of Technology di Pasadena, si distinguono tre «segmenti» neri, uno dei quali è in vicinanza di due profili cellulari, probabili terminazioni di cellule orizzontali. Si tratta di nastri presinaptici in sezione trasversale, il cui aspetto giustifica il nome di «sbarrette», attribuito nel 1953 da Fritiof Sjöstrand. Il citoplasma del cono è occupato molto fittamente da vescicole sinaptiche. I profili cellulari in basso sono sezioni di cellule orizzontali e bipolari con cui il fotorecettore entra in contatto sinaptico.

Nei mammiferi i fotorecettori appartengono a due classi ben distinte sia morfologicamente, sia funzionalmente: i coni e i bastoncelli. Il nome rimanda, sebbene in via troppo semplificatoria, alla loro forma geometrica: coni e cilindri. I coni sono responsabili della visione cromatica ad alte intensità luminose e del riconoscimento dei dettagli fini dell'immagine. I bastoncelli, invece, sono attivi a basse intensità luminose, non contribuiscono alla visione cromatica e in genere forniscono scarse informazioni sulla struttura fine del mondo visivo. Nell'uomo i coni sono strettamente concentrati in un'area molto ben definita e piccolissima della retina, detta fovea. Il restante della retina è coperto in larghissima maggioranza dai bastoncelli. Coni e bastoncelli sono le ultime strutture nervose a ricevere la luce: un raggio luminoso deve, infatti, attraversare tutta la retina prima di raggiungerli e provocarvi tutto quel complesso ordinato e fulmineo di reazioni chimiche e interazioni fisiche che danno l'avvio necessario al riconoscimento visivo del mondo esterno.

Con l'assorbimento di un quanto di luce, o fotone, a livello di una molecola di pigmento specializzata, si inizia una catena di eventi bioelettrici che determinano una serie di microsegnali distribuiti nel tempo e percorrenti le fibre del nervo ottico. Tra l'arrivo di un fotone e l'uscita di un microsegnale dalla retina, sotto forma di microimpulsi elettrici, si intrecciano molti milioni di «operazioni», atte a dare una codificazione sufficientemente dettagliata del tipo di distribuzione della luce in arrivo sui recettori. In questo caso particolare ci soffermiamo sulla prima stazione di trasferimento dell'informazione dal recettore a un diverso tipo cellulare, la cellula orizzontale.

In questa serie di nove fotografie, ottenute al terminale grafico di un calcolatore, viene rappresentata una porzione limitata della superficie di una terminazione di un fotorecettore: un cono di retina di tartaruga. La superficie della membrana di tale cellula appare colorata convenzionalmente in rosso; le regioni in verde e in blu, che compaiono in ogni fotografia, sono le terminazioni di due cellule orizzontali che si invaginano profondamente in una serie di «canyon», formati da pliche della membrana. I nove fotogrammi corrispondono a nove stadi di rotazione dell'immagine, ottenuti allo scopo di percepire l'effetto di profondità dovuto alla complessa struttura tridimensionale della membrana cellulare. Essi sono stati prodotti presso il Computing Center del California Institute of Technology dal gruppo che lavorava in elettrofisiologia e anatomia della visione, e sono stati montati in un'immagine unica dall'autore, in collaborazione con T. E. Ogden e M. C. Citron.

Le cellule orizzontali rappresentano il secondo livello anatomico-architettonico della retina e le loro ramificazioni (dette dendriti dal greco *déndron*, albero) intessono il primo strato retinico. Vorrei concentrare la discussione solo su questo livello strutturale: il punto di contatto funzionale e anatomico tra un fotorecettore, sia esso cono o bastoncello, e una o più cellule orizzontali. La zona specializzata dove due cellule nervose vengono a contatto e si passano l'informazione sotto forma di microcorrenti elettriche o molecole si chiama sinapsi. Quindi argomento di questa «storia» è la sinapsi tra un fotorecettore e una cellula orizzontale. E di una vera storia si tratta, nella quale chi scrive è stato coinvolto direttamente da parecchi anni e con alterne fortune. Queste sinapsi erano state tra le prime a essere studiate ai tempi dei primi «safari» condotti con il microscopio elettronico.

Agli inizi degli anni cinquanta c'erano pochi microscopi elettronici nel mondo e solo alcuni erano stati destinati all'indagine su strutture biologiche. Perciò l'opportunità di dare un'occhiata a retine ingrandite 15 000 - 25 000 volte era

Un nastro presinaptico è qui rappresentato in sezione, come appare nella microfotografia elettronica di pagina 96. Esso affonda in una profonda plica della membrana cellulare, a cui è ancorato da una struttura semicircolare, detta «densità arciforme». Le vescicole sinaptiche *(dischetti più grandi)* riversano il loro contenuto *(dischetti più piccoli)* nello spazio extracellulare. Gli elementi postsinaptici sono due cellule orizzontali, con una zona marcata in grigio (ispessimento sinaptico).

molto scarsa e si riduceva a quei laboratori in cui la locale politica d'istituto ammetteva che potessero servire anche per studi di citologia.

Il Karolinska Institut di Stoccolma era una di queste sedi e Fritiof S. Sjöstrand una di quelle persone atte a determinare tutta una politica d'istituto. Nel 1953, negli atti di un congresso di fisica applicata, uscì un resoconto di 15 righe firmato da lui: in esso si descrivevano nei bastoncelli della retina di una cavia delle enigmatiche «sbarrette». Queste sbarrette apparivano densamente nere e si trovavano nel citoplasma del bastoncello molto vicine alla membrana cellulare. Non sfuggì a Sjöstrand un altro dettaglio importante: le sbarrette si trovavano anche molto vicine alla zona sinaptica del bastoncello. Esse erano state così battezzate perché avevano obiettivamente la forma di segmenti rettilinei neri. Il fatto che apparissero nere al microscopio elettronico era un segno che su di esse l'uranio, che si usa per rendere visibili le strutture cellulari, si era depositato in larga quantità (sotto forma di sale), impedendo agli elettroni di passare e lasciando quindi la propria «ombra» sulla lastra fotografica.

Ma Sjöstrand non aveva intuito la forma corretta, tridimensionale, delle sue sbarrette. Esse apparivano così nella sezione sottilissima (meno di un milionesimo di centimetro) esplorata dal fascio di elettroni del microscopio, ma la sezione non rappresenta l'oggetto. Pensate di tagliare con un piano orizzontale un albero al punto in cui si ramifica e di osservare le tracce lasciate sul piano ideale di sezione da tutti i rami, i ramuscoli e il tronco e magari anche dalle foglie intercettate dal taglio: sapreste ricostruire l'albero da questa sezione?

Il problema di Sjöstrand era simile, anche se più semplice, e il nome che scelse per descrivere le strutture da lui scoperte durante il suo fortunato safari indica che aveva sbagliato nella interpretazione tridimensionale degli oggetti.

Che le sbarrette non fossero tali, ma che fossero dei «nastri» venne riconosciuto dallo stesso Sjöstrand poco dopo e quasi simultaneamente a una ricercatrice francese, Marie Carasso, che capì subito che si trattava di nastri e li chiamò, nella sua lingua, *ruban*. Nel linguaggio scientifico corrente essi vengono chiamati con il termine inglese corrispondente: *ribbon*. Noi li chiameremo qui «nastri presinaptici», indicando così la loro forma geometrica e la loro localizzazione cellulare. A partire dalle prime osservazioni i nastri cominciarono a rendersi manifesti non solo nelle retine, ma anche in molti altri tessuti. Per esempio nastri, ed oggetti a essi molto simili, si notano nelle cellule sensoriali dell'orecchio interno, in quelle della linea laterale dei pesci e nelle cellule dell'epifisi. Sembra quindi che queste piccolissime strutture siano sempre presenti nelle cellule nervose specializzate per avviare una risposta a uno stimolo fisico e siano quindi un poco l'impronta digitale dell'interfaccia sensoriale che mette in contatto l'organismo con il mondo. Questo va bene, ma che cosa ci stanno a fare?

Nel 1975, presso il California Institute of Technology (Caltech) di Pasadena, in California, si costituì un gruppo di ricercatori il cui proposito era di analizzare in dettaglio le modalità di funzionamento del sistema visivo degli invertebrati sul piano sia strettamente fisiologico sia comportamentale, oltre che sul piano, molto più astratto, dei modelli di funzionamento che portano a una buona descrizione del mondo esterno. Il campo d'indagine venne poi ampliato allo studio dei sistemi visivi dei vertebrati.

Quando, nel 1976, venni chiamato a collaborare con il gruppo, questo lavorava presso il Centro di calcolo del Caltech sotto la direzione di Gilbert McCann. C'erano già due elettrofisiologi, Tom Ogden e Marc Citron, e un matematico, John Kroeker. Io avrei dovuto essere il «geografo» della spedizione e curare la morfologia. Il mio primo impatto con i fotorecettori fu traumatico: vidi i famosi «nastri» ma, non avendo letto in precedenza una sola riga su di essi, pensai di esserne io lo scopritore. Durante un pranzo, mi capitò di parlarne con una ricercatrice venuta a visitarci, Anna Bunt, che mi invitò a dire quello che ne pensavo e io le descrissi la «mia» scoperta. Mi disse solo: «Sarebbe meglio che lei passasse più tempo in biblioteca e meno al microscopio». Ci andai e scoprii che quei nastri erano già «vecchi» di ventitré anni e che, per di più, la Bunt era l'unica persona che avesse fatto tutto il possibile su di essi dal punto di vista biochimico e interpretativo. Si ricominciava da zero. L'ipotesi della Bunt, molto simile a un'altra, avanzata quasi simultaneamente in Inghilterra da E. G. Gray, è semplice e affascinante. Ma, per comprenderla, occorre dare, dei nastri, un'altra fetta di informazione.

In moltissimi casi due neuroni stabiliscono un accoppiamento funzionale attraverso un passaggio fisico di un certo tipo di molecole da uno all'altro. In particolare, nel citoplasma di un neurone, quello presinaptico, tutta una catena ancora sconosciuta di eventi biochimici e fisici porta alla costruzione di vescicole di circa 50 nanometri di diametro (1 nanometro è uguale a 10^{-6} millimetri). Nella cavità di ciascuna di queste vescicole trovano posto diverse migliaia di molecole di una sostanza particolare, che qui chiamiamo semplicemente «trasmettitore». Un trasmettitore può essere un amminoacido, l'acetilcolina o un'altra piccola molecola. Le vescicole formatesi nel citoplasma della cellula presinaptica stabiliscono, dall'interno, un contatto con la membrana. Quindi, attraverso un meccanismo fisico in cui svolge una parte importante la tensione superficiale, le vescicole si fondono con la membrana mettendo il loro contenuto in contatto diretto con lo spazio extracellulare. Le molecole contenute un tempo nella cavità della vescicola possono così riversarsi all'esterno e venire a trovarsi nello spazio molto ridotto che divide due cellule vicine: la cellula presinaptica dalla postsinaptica. Le molecole libere di diffondere in questo spazio finiscono per venire in contatto con la superficie esterna della membrana postsinaptica e interagiscono con alcuni gruppi molecolari che sono distribuiti su di essa. Questa interazione porta a una modificazione delle proprietà elettriche locali della membrana che diviene, ad esempio, improvvisamente permeabile a certi ioni presenti nello spazio extracellulare e che prima non erano «ammessi» nella cellula. Il passaggio degli ioni determina una ben precisa corrente elettrica che rappresenta il passo finale dell'accoppiamento funzionale tra le due cellule. Una cellula, la presinaptica, ha trasmesso la sua informazione a un'altra cellula, la postsinaptica, inducendovi una debole, ma precisa corrente elettrica. E i nastri?

La Bunt e quanti altri avevano osservato i nastri avevano notato immediatamente che tutt'attorno al perimetro delle «sbarrette» si allineavano moltissime vescicole sinaptiche. Sembrava che il nastro fosse un luogo di aggregazione, di condensazione, delle vescicole. Inoltre si sapeva benissimo che, a livello di quelle sinapsi, i fotorecettori entravano in contatto funzionale con elementi postsinaptici. Per semplicità, all'inizio di questo articolo, ho detto che i fotorecettori entravano in contatto solo con le cellule orizzontali. In realtà, essi stabiliscono contatti sinaptici anche con un'altra classe cellulare: le cellule bipolari. In sintesi ogni recettore è presinaptico e ha, come partner postsinaptici, due tipi cellulari: le cellule bipolari e le cellule orizzontali. Ciò non modifica comunque molto il problema di fondo che è quello, qui, di descrivere i nastri e di avanzare delle ipotesi sul loro funzionamento.

La semplice osservazione dell'aggregazione delle vescicole lungo i nastri si tradusse subito, nella mente della Bunt e di Gray, in una ipotesi di funzionamento. Ipotesi ben sintetizzata dalla Bunt in questi termini: i nastri presinaptici conducono le vescicole sinaptiche in modo ordinato verso la membrana, al modo stesso di un nastro trasportatore. L'ipotesi viene chiamata, infatti, «del nastro trasportatore». Per visualizzarla pensiamo a un nastro fluttuante vicino a una membrana cellulare (naturalmente dalla parte interna). Il nastro, per ragioni che vedremo dopo, si tiene costantemente in contatto con la membrana lungo un suo lato. Per rappresentarlo meglio, immaginiamo che si tratti di un vero e proprio nastro di forma rettangolare, tenuto fermo per un lato breve e liberamente fluttuante nel citoplasma. Ma il citoplasma non è vuoto, anzi è popolato fittamente di vescicole sinaptiche, che si muovono nel citoplasma stesso in modo casuale, urtandosi l'una con l'altra. Le vescicole che venissero

La ricostruzione tridimensionale di un nastro presinaptico, del tipo rappresentato in sezione nella pagina a fronte, mette meglio in rilievo i rapporti che esso stabilisce con una unità della membrana cellulare. Si notano anche qui la «densità arciforme» che fissa il nastro in fondo all'invaginazione della membrana e l'addensamento delle vescicole sinaptiche sulle due facce del nastro stesso: secondo l'ipotesi avanzata da Anna Bunt, questo assumerebbe la funzione di nastro trasportatore.

La digitalizzazione al calcolatore di una microfotografia come quella di pagina 96 permette di individuare facilmente il contorno della membrana cellulare, i profili delle cellule orizzontali e delle cellule bipolari e due nastri presinaptici. L'immagine è stata costruita a partire da una serie di sezioni che interessavano una notevole porzione della retina. Per ottenerla, una penna collegata a un calcolatore ha permesso di seguire gli elementi prescelti nella microfotografia elettronica e di trasmettere gradatamente alla memoria tutte le coordinate che individuavano i punti critici delle poligonali spezzate, in cui sono state tradotte, digitalizzandole, le linee curve e continue dell'originale.

casualmente in contatto con una delle due facce libere del nastro vi rimarrebbero attaccate. A questo punto inizia il viaggio della vescicola sulla superficie del nastro. La vescicola passa da un mondo tridimensionale, il volume citoplasmatico, a un mondo bidimensionale, la superficie del nastro, lungo la quale si mette in moto e procede sino a raggiungere la membrana cellulare. Qui giunta, viene «condotta per mano» a toccare la membrana e praticamente forzata a fondersi con essa. Il nastro altro non farebbe che rendere più efficiente, più ordinato, più continuo, il casuale bombardamento di vescicole che si ha, invece, nelle sinapsi in cui i nastri sono del tutto assenti. In queste sinapsi, infatti, l'arrivo di una vescicola a contatto con la membrana sinaptica dipende dalla struttura statistica della nuvola di vescicole e dalla loro velocità media. Il nastro introdurrebbe un fattore di efficienza nella sinapsi, funzionando da vero e proprio «nastro trasportatore».

Naturalmente nessuno ha mai visto le vescicole rotolare o essere trasportate sulla superficie del nastro stesso. Le fotografie al microscopio elettronico sono un «congelamento» di una situazione e non ammettono la componente dinamica. Una foto dall'elicottero di una folla uscente da uno stadio, se non ci sono elementi che permettano di riconoscere in quale direzione quelle persone stanno camminando, lascia sempre un'incertezza interpretativa. Non si può decidere se le persone stiano entrando o uscendo dallo stadio, né con quale velocità esse si muovano. Lo stesso accade con le vescicole sinaptiche allineate, nella fotografia, lungo i margini della «sbarretta» o, meglio, distribuite sulle due facce del nastro nella realtà tridimensionale. La potenza della ipotesi della Bunt è nella sua semplicità e nell'immagine che la sintetizza.

Come ogni buona idea che tenta di far luce su un fenomeno complesso, l'ipotesi del nastro dovette però affrontare tutta una serie di difficoltà. Un aspetto un poco periferico al quale mi dedicai nello studio dei nastri fu quello di usarli come test in un complesso programma che avevamo sviluppato al Caltech per la visualizzazione di cellule della retina al terminale di un potente calcolatore. La sequenza delle operazioni che portavano all'immagine sul videoterminale era più o meno la seguente: una retina di rana (o di tartaruga) veniva asportata dall'occhio dell'animale, fissata, inclusa e sezionata per il microscopio elettronico. Tutte le sezioni venivano conservate in serie, in modo da avere la collezione completa di una zona piuttosto ampia di retina. In una serie completa, di circa 200 sezioni, poteva essere compresa tutta la porzione presinaptica di un certo numero di recettori. Per fissare le idee, supponiamo che l'attenzione sia stata rivolta ai bastoncelli. Quando si compie un'osservazione al microscopio, si sceglie, ad esempio, una terminazione di bastoncello e la si va poi a cercare, sezione dopo sezione, sino a collezionare tutte le sezioni in cui essa compare. Con la parola «terminazione» intendo indicare la zona presinaptica del bastoncello, quella dove sono presenti e i nastri e le vescicole sinaptiche. Si fotografano, a un adeguato ingrandimento, le terminazioni e, in particolare, la zona dove si trovano i nastri. Si disporrà così, alla fine, di un certo numero di fotografie, le quali verranno tutte digitalizzate. Si tratta, in pratica, di una semplice operazione che viene condotta con una interfaccia specializzata del calcolatore. Si traccia, su ogni fotografia, il contorno della cellula con una penna collegata al calcolatore in modo tale da inviare in una sua memoria le coordinate attraverso cui la penna passa. Si traduce così, punto per punto, l'immagine.

Naturalmente ho proceduto anche alla digitalizzazione delle «sbarrette» che incontravo, sezione dopo sezione. E, in un programma più ambizioso che venne perfezionato più tardi, ho immesso nel calcolatore anche le coordinate di ciascuna vescicola. In questo modo, di tutta la complicatissima immagine di una fotografia al microscopio elettronico, rimaneva nel calcolatore solo l'elemento essenziale di quel contesto: il profilo della cellula e le sbarrette.

A questo punto, l'operazione di immettere nel calcolatore i dati relativi alle fotografie poteva ritenersi conclusa e finiva la parte di interazione grafica con essa. La situazione adesso era mutata nel senso che la memoria del calcolatore, che era piuttosto capace, era occupata parzialmente da tutte le coordinate di tutti i punti relativi a tutte le linee che erano state tracciate servendosi della penna. Naturalmente ogni coppia di coordinate era accompagnata da una terza coordinata che dava il livello a cui quel punto apparteneva. Quindi, essendo ogni punto di tutta la superficie cellulare (ma anche dei nastri tracciati) individuata da una «tripletta» di coordinate, nella memoria del calcolatore esistevano già tutti gli elementi necessari per ricostruire, nello spazio, l'immagine della cellula (e quella dei nastri). Il passo successivo riguardava «soltanto» la manipolazione di queste triplette e cioè la costruzione matematica di una superficie che collegasse tra loro tutti i profili di tutte le sezioni. Si trattava di una parte di programma di notevole complessità, dovendosi affrontare problemi di tipo topologico, cioè dovendosi «decidere» tra la possibilità o meno di certi collegamenti.

La membrana cellulare non è una semplice superficie cilindroide regolare che si può ricostruire e visualizzare semplicemente. In realtà, essa è un complesso susseguirsi di depressioni, digitazioni, eminenze, cavità, invaginazioni, *canyon* e altre stravaganze morfologiche e topologiche. Per semplicità si considerino solo due contorni di due sezioni adiacenti e separate, nello spazio, da una distanza di circa 90 nanometri. Il mio scopo è stato quello di collegarle tra loro in modo da formare una sorta di palizzata chiusa che le connettesse. Un algoritmo, o procedura di calcolo, sviluppato in precedenza all'Università dell'Utah, permetteva questa costruzione spaziale di «palizzate» usando, come elementi di costruzione, dei triangoli. Quindi l'algoritmo di chiusura provvedeva a costruire una palizzata tra sezioni adiacenti. Questo processo era iterativo nel senso che si completava solo al momento in cui l'ultima sezione veniva collegata con la penultima. A questo punto il calcolatore dava il segnale che era stata costruita la serie completa delle palizzate che collegavano tutte le sezioni tra loro: dalla prima all'ultima.

Prima di procedere, si può cercare di intuire la raffinatezza del programma di costruzione che, presentato così, potrebbe semplicemente parere molto passivo, perfino un poco stupido. Per afferrare un dettaglio importante dell'algoritmo basti pensare a un evento molto comune in una cellula: una biforcazione. Se, in una sezione, si ha una sola traccia, quella del ramo, se ne possono avere due nella sezione successiva, corrispondenti ai due rami originatisi da quel ramo principale. Il calcolatore deve decidere quale sezione collegare e con che cosa collegarla. In questo caso, evidentemente, deve collegarla con entrambi i rami. Ma, in un caso, si può avere «terminazione» in un ramo secondario e «ramificazione» nell'altro ramo.

Il numero di tracce rimane allora costante; ne rimangono infatti due. Come decidere a chi esse appartengono? Come collegarle tra loro? Come, in definitiva, costruire l'algoritmo «di appartenenza»? Evidentemente il programma deve avvalersi di tutta una serie di informazioni aggiuntive, inserite nel programma stesso e basate sulla conoscenza «a priori» delle qualità geometriche e topologiche dell'oggetto da ricostruire. In questo, una certa esperienza di microscopia svolge una parte decisiva nel suggerire al programmatore alcune varianti di programma, necessarie a eliminare dubbi e a evitare assurdi topologici.

La mia ricerca era proprio questa. Benché il lavoro di équipe non mi fosse mai piaciuto, in quel particolare contesto scientifico sembrava una cosa naturale e spontanea, proprio come la formazione di una banda di cacciatori nel Paleolitico. Evidentemente la costruzione delle palizzate non esauriva la ricostruzione spaziale, ma ne dava la base architettonica: i passi successivi furono necessari per dare alla superficie costruita una consistenza visiva. Il colore, l'opacità, la riflettanza, le ombre sarebbero state costruite in seguito. L'immagine finita della superficie cellulare era di una grande bellezza, ma sfuggiva alla mia comprensione. Tutte quelle depressioni, digitazioni, perforazioni, ramificazioni erano davvero significative oppure no?

Il punto rimase insoluto. I nastri vennero fuori e bene: campeggiavano sul terminale in brillanti colori arbitrari e se ne stavano lì in attesa di servire a qualche cosa. Rispetto all'intuizione della Bunt non si era fatto praticamente nessun passo avanti, nulla di più si sapeva rispetto alle «sbarrette» del 1953. Si trattava di un fantastico *tour de force* tecnologico, ma di un ben povero contributo scientifico. Bisognava pensare e c'era da vergognarsi. Ogni ora-macchina costava una notevole quantità di dollari e la ricostruzione di un elevato numero di sezioni - dalle 90 alle 100 - durava tutto un fine settimana. Naturalmente quanto più complessa era la superficie cellulare, quanto più contorti erano i nastri, e quanto più numerose erano le vescicole tanto più lungo risultava il tempo di trasferimento dei dati dalla memoria-base del calcolatore ai programmi di visualizzazione. Certo, una cosa si poteva fare subito e la si fece. Si calcolò con grande precisione la superficie di un grande numero di nastri e si ricostruì la loro forma. Si trattava di dati nuovi in quanto i nastri, tranne in rari casi, ristretti a elementi particolarmente picco-

L'immagine rappresenta due nastri ingranditi di cono di retina di rana, ricostruiti completamente al calcolatore (i colori sono ovviamente arbitrari). Essi sono nelle reciproche posizioni che occupavano nella cellula. Il risultato della ricostruzione semiautomatica tridimensionale ha il vantaggio di poter essere ruotato in tutte le direzioni e quindi di permettere di valutare la posizione stereometrica dei nastri. Inoltre, i dati morfologici su questi, accumulati in memoria, fanno conoscere con molta esattezza la loro superficie, nonostante che la forma non sia né planare né regolare.

La terminazione di un cono di retina è stata qui ricostruita completamente al calcolatore e viene osservata dal punto di vista di un fotone che la colpisce. Si può notare l'estrema complessità delle «sculture» della superficie e come la superficie stessa si possa dividere in prima approssimazione in due zone: una liscia e una molto tormentata. La porzione tormentata corrisponde alle zone sinaptiche; quella liscia alle zone in cui il cono è in contatto con le cellule di Muller, cellule di natura gliale che lo isolano parzialmente dall'ambiente. Anche qui il colore è arbitrario.

li o piatti, non erano stati misurati adeguatamente e non esisteva un vero e proprio «bestiario» per loro.

A questo punto di stallo cominciarono a comparire degli articoli molto interessanti e completamente originali, che rivelavano come i nastri risentissero dell'azione della luce. Un gruppo di Montreal, in Canada (M. A. Ali e J. P. Raynauld), e più tardi un gruppo di Ulm, in Germania (J. Wagner), cominciarono a fornire dati sulla forte riduzione, addirittura sulla scomparsa, dei nastri nei fotorecettori di certi pesci d'acqua dolce durante la notte. Secondo costoro, le prime ore della notte vedevano la scomparsa, o perlomeno la drastica riduzione, del numero dei nastri presenti nei coni. Circa un'ora prima dell'alba i nastri «ritornavano» a comparire nelle terminazioni presinaptiche di questi fotorecettori. Tali dati suscitarono, a tutta prima, un certo sconcerto e diffidenza perché venivano a intaccare quell'atteggiamento, un poco statico, che noi alimentiamo nei confronti delle cellule. In genere queste ci appaiono immobilizzate dai processi istologici, tranne che per i rari film girati *in vivo* al microscopio ottico, o durante l'osservazione delle colture in vitro. Tutto ciò favorisce un atteggiamento di «lettura» della cellula come fosse un organo vivo per definizione, ma morfologicamente piuttosto statico. Dati ancora più recenti del gruppo di Ulm dimostrano che brevissime esposizioni alla luce della retina di certi pesci come la carpa portano a un significativo modificarsi delle strutture terminali dei fotorecettori con le quali le cellule orizzontali stabiliscono i loro rapporti sinaptici. E, al tempo stesso, sembrano influenzare il numero dei nastri. Abbiamo quindi a disposizione un dato importante: i nastri si riducono di dimensioni, o addirittura scompaiono, al buio. Il ciclo si direbbe allora felicemente chiuso e l'attività cellulare della retina si ridurrebbe, in genere, nella condizione di oscurità. Ma qui insorge una difficoltà critica che può portare, a tutta prima, l'idea del nastro a una condizione di contraddizione. Se è pur vero che la condizione di assenza di luce induce una certa riduzione dell'attività che si propaga lungo il nervo ottico, ciò non è vero per nulla all'interno della retina e ancora meno vero è per il livello di funzionamento delle sinapsi tra fotorecettori e cellule orizzontali e bipolari. Anzi, vale il contrario: al buio esiste un continuo traffico di molecole di trasmettitore tra fotorecettori e cellule postsinaptiche; quando la luce colpisce i primi, essi cessano di scaricare il loro quantitativo di trasmettitore, cioè le vescicole non scaricano più il loro contenuto molecolare nello spazio extracellulare; il ritorno del buio riapre, invece, i «canali di comunicazione» tra fotorecettori e cellule postsinaptiche (o di second'ordine). Ciò può apparire contraddittorio con una visione intuitiva della retina. La luce sembrerebbe indurre in attività i recettori ed essi, a loro volta, dovrebbero indurre in attività le cellule, alle quali sono funzionalmente connessi.

Il fatto che, al buio, esista una continua liberazione di trasmettitore tra recettori e cellule del second'ordine venne scoperto circa vent'anni fa dal russo Y. Trifonov, un ricercatore dell'Università di Mosca. La scoperta di Trifonov ha solo avuto conferme dal momento della sua pubblicazione e ormai si tratta di un dato acquisito. Benché esista ancora una notevole incertezza sull'identità chimica del trasmettitore, si può comunque accertare il suo flusso in direzione delle cellule orizzontali durante il buio. Questo dato, considerato assieme alle osservazioni della diminuzione di numero (e di superficie) dei nastri, viene a porre un ostacolo critico alla metafora o ipotesi della Bunt secondo cui i nastri funzionerebbero da nastri trasportatori di vescicole sinaptiche. Se, infatti, il nastro conduce le vescicole alla membrana in modo continuo ed efficiente, come si spiega che alla luce, quando i nastri sono grandi, le vescicole non si scaricano all'esterno e, al buio, quando i nastri sono scomparsi, le vescicole raggiungono la membrana con una elevata frequenza?

A questo punto, ho portato una lieve modifica al mio programma di ricerca. Mentre i miei colleghi continuavano l'indagine elettrofisiologica delle proprietà delle cellule orizzontali e lo sviluppo di modelli matematici adeguati a rappresentarle, mi sono dedicato allo studio dei nastri sotto due condizioni sperimentali estreme. Le tartarughe sono state lasciate un giorno intero alla luce o sono state tenute al buio per due giorni. Le loro retine sono state trattate come d'abitudine e, per finire, le immagini dei nastri adattati al buio sono state visualizzate al videoterminale. Non c'era dubbio: il buio, anche nelle tartarughe, riduceva l'area dei nastri di più del 50 per cento. Vedremo poi i particolari di questa riduzione di superficie e cioè in che modo la forma dei nastri andasse o non modificata durante il processo di riduzione o distruzione. Per il momento si poteva pensare a una soluzione opposta all'ipotesi del nastro. I nastri, forse, agivano come «cani da guardia» nel senso di impedire alle vescicole sinaptiche di raggiungere la membrana cellulare quando la luce colpiva i recettori. La loro grande superficie, alla luce, assicurava una notevole efficienza nel bloccaggio delle vescicole. Al buio, invece, essi diminuivano, di superficie e le vescicole erano libere di muoversi sino a raggiungere la membrana. La variante dell'ipotesi del nastro introduceva, per il nastro, una duplice interpretazione postulando due modi di funzionamento: uno caratterizzato dalla luce e l'altro dal buio.

In questa ricostruzione si notano due popolazioni di vescicole sinaptiche *(dischetti)*, ciascuna associata a una faccia (*a* o *b*) di un nastro presinaptico di cono di retina di tartaruga dopo 24 ore ininterrotte di esposizione alla luce. Si può notare come le vescicole siano distribuite in modo disordinato e occupino solo una piccola porzione della superficie del nastro presinaptico.

La possibilità di ricostruire con grande precisione i nastri assieme alle vescicole sinaptiche a essi associate mi ha permesso di osservare come, sulla superficie di un nastro adattato alla luce, le vescicole siano poche e mal distribuite, mentre, nel caso di nastri «tenuti al buio», le vescicole si distribuivano secondo una perfetta simmetria esagonale. Si potrebbe «leggere» questo dato dicendo che, alla luce, è rimasto «attaccato» al nastro solo un certo numero di vescicole mentre, al buio, le vescicole occupano tutta la superficie in modo compatto e omogeneo. L'aggregazione delle vescicole sui nastri è un fatto ben noto, ma essa non è mai stata «trattata» quantitativamente in modo adeguato. Questo compito viene semplificato dalla possibilità, offerta dal calcolatore, di valutare con grande velocità le coordinate di un ampio numero di vescicole nelle regioni circostanti un nastro e, naturalmente, nei due casi di adattamento al buio e alla luce. Si è riusciti, così, a calcolare la distribuzione della densità delle vescicole nelle immediate vicinanze di un grande numero di nastri. In tutti i casi, si è trovato che le vescicole raggiungono la concentrazione massima alla superficie del nastro stesso e questo era il dato scontato; ma la rarefazione massima non coincide con la distanza massima dalla superficie del nastro e questo dato non era atteso. Una sua «lettura» ci permette di indovinare la situazione di una sorta di campo attorno al nastro, un campo la cui natura non è nota, ma che induce nella popolazione di vescicole circostante un'alterazione della densità della loro distribuzione.

Per avere un'idea di come sia tale distribuzione in assenza e in presenza dei nastri si può usare il calcolatore e fargli generare una sorta di mappa geografica in cui la densità di linee stia a indicare la locale concentrazione di vescicole sinaptiche. Si crea così una mappa del tipo di quelle basate sulle isocline, in cui le pendici delle montagne vengono simboleggiate dalla densità delle linee che passano per quel punto. Pareti ripide sono indicate da linee molto fitte, pendici quasi pianeggianti da linee molto spaziate. Nella nostra codificazione le zone ad alta densità sono rappresentate da alta densità di contorni e vale il reciproco per le basse densità. Le due mappe, messe a confronto, descrivono molto bene il concentrarsi di vescicole nelle vicinanze dei nastri e, inoltre, dimostrano la struttura «disordinata» delle vescicole lontano da essi.

Quale sia la natura del «campo» lo si può dedurre, molto indirettamente, da un altro dato morfologico concernente le vescicole. Nel caso in cui esse raggiungano la massima densità, e quindi il massimo ordine, sulla superficie di un nastro «tenuto» al buio, esse si distribuiscono con la ben nota simmetria esagonale che rappresenta il modo «migliore» di occupare un piano da parte di una popolazione di sfere a diametro uniforme. Ma, se si considera quante sferette delle dimensioni delle vescicole sinaptiche possono essere contenute su una data superficie, si osserva che questo dato calcolato è molto inferiore a quello osservato. Nel caso del nastro «al buio» le vescicole si distribuiscono in modo regolare secondo la simmetria di un alveare, ma il valore della loro densità è inferiore a quello previsto dalla condizione teorica. Tale dislivello dipende dal fatto che la popolazione teorica si presenta in condizioni di reciproca tangenza, ossia ogni sferetta è in contatto di tangenza con altre sei sferette. Nel caso del nastro «al buio», invece, ogni vescicola, pur trovandosi al centro di una rosa di altre sei, non le tocca mai. Attorno a ognuna di esse è definibile un'area di nessuno, in cui nessun'altra vescicola viene ammessa.

Un dato in appoggio a quanto misurato sulle microfotografie di nastri ricoperti di vescicole viene da uno studio di un gruppo di ricercatori giapponesi, che ha misurato un analogo valore di «raggio di repulsione» attorno a vescicole sinaptiche tratte da un'altra sinapsi. Essi hanno misurato la distanza intervescicolare in popolazioni libere di vescicole sinaptiche e hanno raggiunto la conclusione che, attorno a ogni vescicola, si possa individuare un campo di tipo elettrostatico. Questo campo coulombiano ricorda molto quello che si genera attorno a una sferetta metallica carica, cioè tale per cui le cariche distribuite sulla sua superficie non abbiano la possibilità di lasciarla. Se poniamo due di tali sferette cariche di elettricità dello stesso segno vicine l'una all'altra, esse saranno sottoposte a un reciproco campo repulsivo che le porta ad allontanarsi tra loro. Esiste naturalmente una «ragione sufficiente» per un tale campo repulsivo tra vescicole sinaptiche: queste sono sostanzialmente costituite da membrane la cui natura è identica in tutte. Di conseguenza, se la membrana delle vescicole non è elettricamente neutra (e si sa che non lo è), allora ne risulta che i gruppi carichi delle proteine di membrana, o le teste polari dei lipidi, che contribuiscono a formare la parete vescicolare, devono creare un campo elettrico sempre dello stesso segno. Due vescicole di questo tipo, quindi, non possono avvicinarsi troppo tra loro.

Il problema derivante dal fatto che le vescicole sono sostanzialmente immerse

Il grafico in basso rappresenta il variare della concentrazione delle vescicole sinaptiche nello spazio attorno a 70 nastri presinaptici. Lo schema in alto mostra come sono state ottenute le misure: si contavano quante vescicole erano contenute nei quadrati di un foglio trasparente sovrapposto alla microfotografia elettronica. Si rileva come la concentrazione raggiunga il massimo a livello della superficie del nastro e come essa diminuisca sino a raggiungere un valore minimo alla distanza di 0,3 micrometri dalla superficie. Oltre questo valore minimo, la concentrazione risale sino ad assumere valori stazionari. Questi ultimi caratterizzano la concentrazione media delle vescicole sinaptiche in zone che non sono sotto il controllo dei nastri presinaptici.

Una mappa costruita al calcolatore *(in alto)* informa sulla concentrazione delle vescicole sinaptiche in una terminazione di cono. In essa si osserva che i nastri sono assenti e il valore medio della densità delle vescicole sinaptiche è basso ed è stabilizzato, pur oscillando attorno a un valore medio abbastanza costante. La mappa qui sopra, invece, mostra una terminazione di cono in cui sono presenti due nastri: si osserva attorno ad essi un notevole accumularsi di vescicole, simboleggiato dall'infittirsi delle isocline. La figura sottolinea l'effetto di «campo» associato ai nastri presinaptici e l'estensione in cui esso produce conseguenze misurabili sulla «nube» costituita da vescicole sinaptiche. Queste immagini sono state ottenute dall'autore in collaborazione con M. C. Citron presso il Computing Center del California Institute of Technology di Pasadena.

in acqua non altera molto i dati del problema in quanto esse risentono la presenza una dell'altra attraverso una complessa ridistribuzione delle molecole d'acqua interposte. Il dato essenziale è che le vescicole sono associate a un campo di tipo negativo in quanto i gruppi carichi delle molecole che costituiscono la membrana sono negativi. Con questo abbiamo «spiegato» la relativa bassa densità delle vescicole sulla superficie del nastro, ma non abbiamo spiegato come mai esse risentano del «campo» d'attrazione del nastro. Un elemento, anch'esso derivante dallo studio biochimico della Bunt, può aiutare a dare un'ipotetica linea di spiegazione di questo effetto.

I nastri vengono «digeriti» con molta efficienza da enzimi che appartengono alla classe generale delle pronasi, enzimi proteolitici che smantellano le molecole proteiche. Fotorecettori incubati in presenza di pronasi lasciano intravvedere solo pallide ombre laddove in precedenza c'erano i nastri. Ora, le pronasi attaccano con una certa specificità i residui dell'arginina e della lisina, due amminoacidi i cui residui hanno cariche positive. In prima approssimazione si può pensare che un nastro si presenti, per una vescicola, come un oggetto planare a carica nettamente positiva. Ne deriva, di conseguenza, un campo attrattivo per le vescicole che abbiamo visto essere dotate di cariche negative.

In effetti, durante l'embriogenesi, quando i nastri si formano nei recettori in via di sviluppo, si assiste quasi immediatamente anche al formarsi del complesso vescicole-nastro. Ma, se il nastro «vuoto» era sede di un campo attrattivo per le vescicole che gli galleggiavano attorno, una volta ricoperto da vescicole negative esso diveniva un oggetto fortemente negativo per la restante popolazione di vescicole, situate a una certa distanza. Questo fatto darebbe una spiegazione assai semplice (e forse semplicistica) della zona di rarefazione che si incontra a circa 0,3 micrometri di distanza dalla superficie del nastro. Supponiamo che una vescicola si trovi, rispetto al complesso vescicole-nastro, a una distanza tale da non risentirne gli effetti repulsivi. Essa si muoverà con una qualche strategia di tipo browniano, determinata dagli urti reciproci tra vescicole. Se, però, per effetto della «passeggiata casuale in tre dimensioni», si verrà a trovare troppo vicina al complesso vescicole-nastro, essa sarà respinta dal suo potente campo elettrostatico negativo e quindi si verrà a individuare, attorno al nastro, una zona a bassa concentrazione vescicolare: la zona di rarefazione che abbiamo incontrato, appunto, tutte le volte che abbiamo misurato la densità delle vescicole nel volume adiacente al nastro stesso.

A questo punto rimane il problema del moto vero e proprio delle vescicole che sono «poggiate» sulla superficie del nastro. Sappiamo che esse devono, prima o poi, raggiungere la membrana cellulare e quindi devono muoversi rispetto al nastro. Di conseguenza deve instaurarsi, sulla superficie del nastro, un moto di scorrimento di tali sferette. Chi fornisce l'energia? Il nastro presenta, secondo questa ipotesi, una superficie allo stesso potenziale elettrico. O, come si dice, una superficie equipotenziale. Un esempio analogo è dato dalla superficie terrestre, che è isopotenziale nei confronti dell'accelerazione di gravità. Lo spostamento di una massa, parallelamente alla superficie terrestre, deve avvenire, dunque, senza lavoro in quanto il punto di applicazione della forza si muove ortogonalmente rispetto alle linee di forza del campo stesso. Ne deriva che il moto delle vescicole è vincolato, nel senso che esse non possono lasciare la superficie del nastro, ma sono relativamente libere di scorrervi sopra. Il lavoro che devono fare per muoversi deve essere ristretto a quello che serve a controbattere i vari coefficienti di attrito, i quali, senza dubbio, si generano tra vescicole e alone di molecole d'acqua orientate, che si portano appresso. Ma, trattandosi di basse velocità, è probabile che le forze di attrito da vincere non siano molto grandi è che solo una piccola porzione di energia sia sufficiente per tenerle in moto. Ciononostante dobbiamo ricercare una possibile fonte di energia che spieghi il moto delle vescicole, ma soprattutto che ne spieghi il «ricambio». Infatti, al buio, i nastri sono sempre pieni di vescicole nonostante il continuo «sgocciolamento» di queste verso la membrana. Ciò indica che i vuoti provocati dalla perdita di vescicole vengono immediatamente sostituiti da nuove vescicole, anche se - come abbiamo detto in precedenza - a causa del forte campo negativo repulsivo le vescicole hanno difficoltà a raggiungere la superficie del nastro o, addirittura, ne sono del tutto impedite.

Come si spiega allora il pronto riempirsi dei vuoti provocati dal continuo eliminarsi di vescicole? Se riprendiamo l'analogia con il nastro carico sospeso nell'acqua e ci domandiamo in che modo una vescicola, carica dello stesso segno, possa atterrarvi, ci accorgiamo che l'elettrologia viene in nostro aiuto. Esiste, infatti, un punto di atterraggio «autorizzato», una specie di tallone di Achille, del nastro. Sappiamo che i nastri hanno una forma approssimativamente rettangolare e che sono «ancorati» per un lato lungo alla parete interna della membrana cellulare. Da questo deriva che esiste tutto un lungo perimetro liberamente fluttuante all'interno della cellula. Se una vescicola si avvicina al nastro secondo una traiettoria ortogonale alla sua superficie, essa viene respinta da una forza elettrostatica massima, in quanto ogni elemento della superficie contribuisce al massimo al campo e il valore della forza repulsiva è legato alla somma elementare di tutti i singoli contributi. Ma se la vescicola, per caso, si avvicina di taglio al nastro, allora il campo repulsivo è praticamente nullo in quanto la superficie è vista sotto un angolo minimo.

Esisterebbe quindi, per le vescicole, un punto preferenziale di atterraggio e si tratterebbe del perimetro interno del nastro. Si dimostra che la zona di rarefazione, che è massima in direzione ortogonale alla superficie del nastro, è minima nella direzione a essa parallela. In molte fotografie si osservano «code» di vescicole che sembrano in attesa di essere ammesse a toccare la superficie del nastro. I progressivi urti delle nuove vescicole potrebbero fornire la poca energia necessaria a tenere in moto tutta la popolazione che ormai scorre su tale superficie.

Rimangono altri problemi, ma uno è centrale: come può accadere che le vescicole siano libere di scorrere sulla superficie del nastro al buio e, invece, vi si «appiccichino» alla luce? Se, infatti, è corretta, la variante che vede nei nastri dei «cani da guardia» per le vescicole, occorre spiegare questo secondo meccanismo (o perlomeno avanzare un'ipotesi su di esso). Le vescicole sinaptiche, nelle sinapsi tra nervo e muscolo, sono caratterizzate dal possedere quello che si chiama «il punto del calcio». Si tratta di una zona specializzata della membrana vescicolare che deve orientarsi in un modo particolare rispetto alla membrana cellulare perché avvenga la fusione con essa. Questo allineamento è mediato dagli ione Ca^{++}. Quando questi sono presenti, allora le vescicole possono attaccarsi alla membrana, fondersi e scaricare il loro carico molecolare all'esterno.

Non esistono prove che le «nostre» vescicole posseggano anch'esse il «punto del calcio», ma esse non sembrano comportarsi differentemente dalle altre e il meccanismo della loro fusione con la membrana cellulare è del tutto simile a

Le stesse mappe della pagina a fronte vengono qui rappresentate secondo una «prospettiva a volo d'uccello». In assenza di nastri presinaptici *(in alto)*, le vescicole presinaptiche sono distribuite con una certa regolarità, mentre in presenza di nastri *(qui sopra)* si osservano zone con una notevole intensità vescicolare. Nel testo viene delineata una semplice ipotesi che attribuisce tali effetti allo stabilirsi di un campo elettrostatico reciprocamente attrattivo tra vescicole e nastri. Si sa, infatti, che le vescicole sinaptiche possono essere considerate sferette cariche negativamente e ci sono indizi biochimici che i nastri possano essere associati con notevoli cariche superficiali di segno positivo. Queste figure sono state ottenute dall'autore in collaborazione con M. C. Citron presso il Computing Center del California Institute of Technology di Pasadena.

quello incontrato in altre sinapsi. Potrebbe accadere che, nel nastro «illuminato», si sviluppi per mezzo dello ione calcio un ponte che lega stabilmente le vescicole alla sua superficie. Ma quest'ipotesi riposa su una disponibilità di Ca^{++} durante l'illuminazione dei fotorecettori. Esiste tutta una serie di osservazioni che tendono a far ritenere che l'iniezione di Ca^{++} in un recettore induca in esso dei comportamenti elettrici identici a quelli che si ottengono illuminandolo. Inoltre, esistono prove che il processo di trasferimento dell'energia del fotone alla cascata di reazioni biochimiche che porta alla risposta elettrica del fotorecettore include un intervento attivo del calcio. Lo studio dei fenomeni di contrattilità in muscoli, flagelli e ciglia dimostra la funzione di accoppiamento dinamico tra ioni calcio e molecole proteiche. In via ipotetica, non è illogico pensare a un aumento della concentrazione intracellulare degli ioni calcio durante l'illuminazione del recettore; al buio il vincolo tra vescicole e nastro si scioglie ed esse ritornano libere di scorrere sulla superficie, di raggiungere la membrana cellulare e di scaricare il loro contenuto nello spazio sinaptico.

È proprio questo un punto critico di tutta la struttura vescicole-nastro: il suo funzionamento al buio. Vediamo in che modo questa «mediazione geometrica» indotta dal nastro può portare a importanti conseguenze sul processo stesso della visione al buio. Sappiamo che la luce blocca il passaggio del trasmettitore dal fotorecettore alle cellule orizzontali e bipolari e che il buio lo riattiva. Questo vuol dire molto semplicemente che, se si riesce a bloccare il passaggio del trasmettitore sinaptico tra fotorecettore e cellula orizzontale, per questa è come se fosse arrivato un lampo di luce. La cellula orizzontale, di per sé, non ha modo di controllare se la luce è davvero arrivata o no. Il processo di trasduzione dell'energia del fotone in modificazione di strutture molecolari è tutto a carico del recettore, che ne informa la cellula orizzontale bloccando il flusso di trasmettitore. Quindi se, per accidente, cessa il flusso di trasmettitore dal fotorecettore mentre la retina è tenuta al buio, la cellula orizzontale deve segnalare alle altre che sono a essa connesse che il buio è stato interrotto da un breve lampo di luce.

Quest'informazione periferica falsa viene poi inviata a livelli più complessi, dove viene elaborata secondo schemi ipotetici, ma che ormai non hanno più nessun possibile controllo sull'avvenuta o meno stimolazione luminosa del fotorecettore. I nastri possono ovviare a questo inconveniente assicurando un continuo regolare «sgocciolamento» di vescicole. In effetti, le vescicole devono atterrare sul nastro secondo la strategia «disordinata» della nube statistica a cui appartengono e da cui si separano. Ma, lungo il cammino sul nastro, raggiungono una compattezza e un ordine, dimostrati dalle fotografie al microscopio elettronico e tali per cui, all'arrivo, esse sono sistemate con grande regolarità.

Per tutti i nastri ricostruiti al calcolatore e misurati, si è constatato che erano elementi molto vicini al rettangolo e con un rapporto di 1:3 tra i due lati. Quando, al buio, si riducevano fortemente di dimensioni, scendendo a meno del 50 per cento della loro superficie, i rapporti tra i lati venivano mantenuti. Esiste evidentemente un sistema di controllo che assicura, sia durante la sintesi dei nastri sia durante la loro distruzione enzimatica, una costanza di rapporto geometrico tra lati lunghi e lati brevi. Il cammino percorso, ipoteticamente, dalle vescicole è quello parallelo ai lati brevi, che devono essere però sufficientemente estesi da assicurare un ordine, derivante dallo scorrimento delle vescicole, ma non troppo estesi per non creare un blocco del movimento dovuto all'aumentare delle forze di attrito. Infatti la distribuzione delle lunghezze dei lati brevi è molto omogenea.

Possiamo adesso aggiungere un'ultima tessera al mosaico misto di ipotesi, dati, esperimenti condotti sui nastri. Essa coinvolge un concetto un poco complesso, ma molto affascinante: il «rumore sinaptico». Come abbiamo veduto, la cellula presinaptica scarica molecole di trasmettitore nello spazio sinaptico. E queste molecole vengono in contatto con la membrana della cellula postsinaptica. Il loro arrivo è segnalato da piccole correnti locali, derivanti da locali eventi di apertura e di chiusura di canali, che fanno passare o impediscono il passaggio agli ioni presenti rispettivamente nel citoplasma della cellula postsinaptica e nello spazio sinaptico stesso. Se si inserisse, in vicinanza della membrana postsinaptica, un microelettrodo capace di registrare (in virtù delle sue caratteristiche di conduzione elettrica e dell'elettronica a cui è associato) si riuscirebbe a documentare una serie di piccoli «eventi», o microcorrenti, che si susseguono nel tempo. La distribuzione temporale di questi microeventi, detti anche potenziali-miniatura, riflette naturalmente la distribuzione temporale dei successivi arrivi delle molecole di trasmettitore. E queste, a loro volta, dipenderanno naturalmente dai ritmi seguiti dalla cellula presinaptica nel riversarle nello spazio sinaptico.

I potenziali-miniatura si possono misurare molto bene nelle cellule bipolari. Nelle cellule orizzontali tale misura delicata è impossibile per il fatto che tutte le cellule di questa classe sono strettamente vincolate tra loro da punti di contatto a bassa resistenza elettrica. La conseguenza dell'esistenza di una tale rete è che piccoli segnali intracellulari vanno subito dispersi su grandi superfici. Immaginate, per esempio, di «iniettare» una piccolissima corrente istantanea in una rete di elementi che conducano bene l'elettricità. In tempi brevissimi questo «impulso elettrico» si distribuirà in un grande numero di elementi. Nelle cellule bipolari, invece, i segnali piccoli rimangono segregati a causa del relativo isolamento elettrico che caratterizza questa classe di cellule. (Ricordiamo che le cellule bipolari sono anch'esse postsinaptiche rispetto ai fotorecettori.) Se in una cellula bipolare di uno squalo si registra il rumore sinaptico, si nota che esso ha una struttura statistica molto differente nelle due condizioni di luce e di buio. Due ricercatori inglesi, A. H. Ashmore e G. Falck, sono riusciti a eseguire queste delicatissime misurazioni e hanno dimostrato che il rumore sinaptico «al buio» presenta due componenti statistiche di cui una scompare quando la retina viene illuminata.

Che cosa si intende con «componenti statistiche di un rumore»? Una di esse è rappresentabile da un arrivo di numerosi eventi molto vicini nel tempo, ma tutti più o meno simili tra loro. L'altra da arrivi molto più spaziati nel tempo, ma appartenenti a diversi livelli di energia. La prima, quella ad alta frequenza, scompare quando la retina viene illuminata. La seconda, quella a bassa frequenza, rimane anche con l'illuminazione.

Ora, le cellule bipolari presentano due tipi diversi di sinapsi con i fotorecettori. Un tipo è la sinapsi che abbiamo veduto sinora: una sinapsi controllata da un nastro. L'altro, detto «sinapsi piatta», non presenta invece nessun nastro. In esso, le vescicole sono libere di urtare la membrana a seconda della loro distribuzione statistica. Il fatto che, alla luce, la componente ad alta frequenza (che è anche quella «meno disordinata») scompaia e rimanga invece quella a bassa frequenza può essere spiegato con l'ipotesi avanzata prima sul funzionamento dei nastri: alla luce i nastri bloccano le vescicole e quindi solo le sinapsi piatte funzionano (una sola componente di rumore è presente); al buio i nastri cominciano a funzionare assieme alle sinapsi piatte (due componenti di rumore sono simultaneamente presenti). Di recente, l'anatomista francese R. F. Dacheux, ha dimostrato che la stessa cellula bipolare può presentare entrambi i tipi di sinapsi.

Mi piace finire la storia dei nastri presinaptici proprio in Francia, dove essi erano stati battezzati con il loro bel nome di *ruban*. Per il resto, infatti, al momento in cui si poteva passare a un loro studio più incisivo, usando una sorta di tracciante biochimico, la perossidasi, che ci avrebbe permesso di seguire il moto delle vescicole e quindi di dedurre la loro «temperatura termodinamica», il gruppo del Caltech in cui lavoravo venne sciolto, essendo state respinte le richieste di finanziamento. Nel giro di tre mesi l'apparecchiatura fu venduta, Marc Citron se ne andò al Children Hospital di Los Angeles, Tom Ogden ritornò a fare l'oftalmologo alla University of Southern California, mentre di John, il matematico, si sono perse le tracce. Il direttore del gruppo, Gilbert McCann, è ora a godersi i suoi cavalli arabi nel ranch che possiede nell'alta California. Io, per guadagnarmi da vivere, ho fatto per ancora un po' di tempo il tecnico istologo al Caltech. Là, dove avevamo lavorato per tre anni, nuova gente, nuove macchine, nuovi problemi. Una lezione assai salutare.

TEMI METAMAGICI

di Douglas R. Hofstadter

Il calcolo della cooperazione è messo alla prova con un nuovo dilemma

Un bel giorno ricevete una lettera da S. N. Platonia, il ben noto petroliere plurimiliardario, in cui vi si annuncia che 20 persone, conosciute per le loro capacità razionali, sono state scelte per partecipare a un piccolo gioco. «Lei - dice la lettera - è una di queste persone. Ognuno di voi ha la possibilità di vincere un miliardo di dollari, messo in palio dall'Istituto Platonia per lo Studio dell'Irrazionalità Umana. Ecco le modalità di partecipazione. Se lo desidera, può inviare un telegramma con il suo nome all'Istituto Platonia, Frogville, Oklahoma. Può far addebitare la spesa all'Istituto. Se risponde entro 48 ore, il miliardo è suo - a meno che ci siano due o più risposte, nel qual caso il premio non è attribuito a nessuno. Se non vi sarà alcuna risposta, non verrà dato nulla a nessuno.»

Non avete alcun modo per sapere chi sono gli altri 19 partecipanti; la lettera di Platonia, anzi, afferma che l'offerta sarà ritirata nel caso si scopra che un partecipante ha tentato di conoscere l'identità o di stabilire qualche contatto con un altro partecipante. Per di più, l'eventuale vincitore deve sottoscrivere l'impegno a non dividere mai il premio con altri partecipanti. Questo per reprimere ogni idea di cooperazione, sia prima che dopo l'attribuzione del premio.

Il fatto brutale è che nessuno saprà che cosa fanno gli altri. Chiaramente ognuno vuole quel miliardo e altrettanto chiaramente si rende conto che se *non* invia il proprio nome non avrà alcuna possibilità. Questo significa che a Frogville arriveranno 20 telegrammi, dimostrando così che anche ai vostri eccezionali livelli di razionalità non vi è alcun modo di risolvere una situazione tanto aggrovigliata e tormentosa?

Questo è il Dilemma di Platonia, una piccola sceneggiatura che ho elaborato recentemente per affrontare in modo migliore il Dilemma del prigioniero, di cui ho parlato il mese scorso. Il Dilemma del prigioniero può essere formulato in termini simili a questi. Immaginate di ricevere dall'Istituto Platonia una lettera in cui vi si annuncia che siete stati scelti insieme a *un* altro pensatore razionale per concorrere a un modesto premio in denaro. Come prima, a entrambi viene chiesto di rispondere entro 48 ore all'Istituto Platonia con un telegramma a carico del destinatario. Il telegramma deve accludere, oltre al vostro nome, solo la parola «coopero» o la parola «defeziono». Se entrambi telegrafate «coopero», riceverete 3 dollari a testa. Se entrambi telegrafate «defeziono», riceverete 1 dollaro a testa. Se inviate parole diverse, chi coopera non riceverà nulla, mentre chi defeziona riceverà 5 dollari.

Che cosa scegliereste? Sarebbe bello se entrambi cooperaste, in modo da ricevere tutti e due 3 dollari, ma questo non sembra un po' improbabile? Dopo tutto, chi ha voglia di farsi fregare da un odioso che si prende 5 dollari per la propria vile defezione? Certo non *voi*! Perciò decidereste probabilmente di non cooperare: una scelta riprovevole ma apparentemente necessaria. Naturalmente, entrambi ragionate allo stesso modo e arrivate alla stessa conclusione. Quindi entrambi defezionerete e vi prenderete un solo dollaro a testa. Se invece entrambi aveste voluto correre un rischio, avreste avuto 3 dollari a testa. Peccato!

Il disagio che viene spontaneo provare davanti a questa analisi apparentemente logica del Dilemma del prigioniero in versione «colpo secco», mi portò a inviare la seguente lettera a 20 amici:

«Caro ———;
invio questa raccomandata a 20 di voi (amici miei sparsi in tutto il paese). Vi propongo una partita al Dilemma del prigioniero in versione "colpo secco" con pagamento in denaro (fornito da "Scientific American"). La cosa è molto semplice; ecco come funziona.

«Ognuno di voi mi deve scrivere semplicemente una lettera: *C* o *D*, rispettivamente per "coopero" o "defeziono". La lettera sarà la tua mossa in un Dilemma del prigioniero con *ciascuno* degli altri 19 giocatori. La matrice di pagamento usata è l'allegato numero 1 *(si veda l'illustrazione in questa pagina)*.

«Se tutti inviano *C*, tutti vincono 57 dollari, mentre se tutti inviano *D*, tutti vincono 19 dollari. Non si può perdere! Naturalmente, chi invia *D* avrà almeno quanto ogni altro. Se, per esempio, 11 persone inviano *C* e nove *D*, gli 11 *C* riceveranno 3 dollari a testa da ciascuno degli altri *C* (quindi 30 dollari) e non riceveranno nulla dai *D*. Tutti i *C*, quindi, riceveranno 30 dollari. I *D*, invece, avranno 5 dollari a testa da ciascuno dei *C* (55 dollari) e 1 dollaro da ciascuno degli altri *D* (8 dollari), per un totale di 63 dollari. Qualsiasi sia la distribuzione, ai *D* andrà sempre meglio dei *C*. Naturalmente, più *C* ci sono, meglio sarà per *tutti*!

«Sia chiaro che nel fare la tua scelta non devi pensare a *vincere*, ma semplicemente a ricevere più *denaro* possibile. È quindi meglio per te ricevere 30 dollari (inviando *C* insieme ad altri 10, anche se i nove *D* prendono più di te), piuttosto che ricevere 19 dollari (dicendo *D* insieme a tutti gli altri, in modo che nessuno possa "batterti"). Inoltre, non devi pensare di incontrarti in seguito con altri partecipanti e dividere con loro il guadagno: il vostro scopo non è quello di massimizzare il numero totale di dollari sborsati da «Scientific American», ma solo quello di massimizzare il numero di dollari che vengono a *voi*!

«Ovviamente, la tua speranza è quella di essere l'unico che defeziona, in modo da sbancare tutto: con 19 *C*, tu prenderesti 95 dollari e loro 18 volte 3 dollari, cioè 54 dollari. Ma perché sto qui a fare molti plicazioni ed esempi? Tu hai un'intelligenza brillante; e anche gli altri. Tutti egualmente brillante, direi. Tutto quel che devi fare, quindi, è dirmi la tua scelta. Voglio tutte le risposte per telefono (chiamate a carico del destinatario) *il giorno stesso in cui ricevete la lettera.*

«Mi raccomando (è *quasi* sottinteso, ma non del tutto) di non cercare di consultarti con altri che pensi siano stati invitati a partecipare. Anzi, ti prego di non consultarti proprio con nessuno: lo scopo è quello di vedere quello che farete di vostra iniziativa, isolati. Insieme alla scelta, infine, gradirei una breve spiegazione del *perché* hai fatto quella particolare scelta.

Tuo
Doug H.

«P.S. Incidentalmente, ti può essere utile immaginare una situazione analoga, con l'unica differenza, rispetto a questa, che tutti gli altri giocatori hanno *già* fatto la loro scelta (diciamo una settimana fa) e quindi tu sei l'ultimo. Cosa fai a questo punto? Rispondi *D*, sapendo bene che le *loro* risposte sono già depositate? Supponi ora che immediatamente dopo aver detto il tuo *D* (o il tuo *C*) ti venga detto che in realtà gli altri *non hanno* depositato davvero le loro risposte ma che lo faranno tutti oggi. Ritireresti la tua risposta? E se tu sapessi (o almeno ti venisse detto) che sei la *prima* persona a cui è stata richiesta una risposta?

«Un ultimo punto su cui riflettere: che cosa faresti se la matrice di pagamento fosse come quella dell'allegato 2?» *(si veda l'illustrazione in questa pagina)*.

		Giocatore *B*	
		Coopera	Defeziona
Giocatore *A*	Coopera	(3,3)	(0,5)
	Defeziona	(5,0)	(1,1)

Dove (*x*, *y*) significa che il Giocatore *A* guadagna *x* dollari e il Giocatore *B* guadagna *y* dollari.

Una matrice di pagamento per il Dilemma del prigioniero

Voglio sottolineare che questa situazione non è un Dilemma del prigioniero *iterato* (di cui ho parlato nell'articolo del mese scorso). È un Dilemma del prigioniero a «colpo secco», con molti partecipanti. Non è possibile apprendere qualcosa nel tempo sulle tendenze di gioco degli altri. Non sono applicabili, pertanto, le «lezioni» descritte il mese scorso, dal momento che dipendono dall'iterazione della situazione. Tutto quel che potevano pensare i mittenti della mia lettera era questo: ci sono 19 persone in qualche modo simili a me, tutte nella stessa barca, tutte alle prese con lo stesso problema. In altri termini, non c'è nulla a cui affidarsi al di fuori della pura ragione.

Mi divertii molto nel preparare la lettera, decidere a chi mandarla, prevedere le risposte e poi riceverle. Prima di rivelarvi le risposte, vi invito a pensare a come vi sareste comportati in quella medesima situazione. Vorrei soprattutto che prendeste in seria considerazione l'affermazione che tutti i partecipanti hanno un'intelligenza molto brillante. Lasciate anzi che mi dilunghi un attimo su questo argomento, perché penso che le persone a cui mi ero rivolto non abbiano afferrato in pieno quello che intendevo. Immaginate quindi che la lettera contenga questo paragrafo chiarificatorio:

«Siete tutti persone molto razionali. Quindi, non c'è quasi bisogno di dirti che devi fare quella che consideri la scelta più razionale: in particolare, devi mettere da parte ogni idea di moralità o colpa, ogni senso di vago malessere. Alla base della tua decisione deve stare solo il ragionamento (compreso, naturalmente, il ragionamento sul ragionamento degli altri). E ti prego di ricordare che a tutti è detto questo (compreso *questo*!)!».

Speravo - e mi aspettavo - un esito particolare dell'esperimento. Ricevute le risposte, nei giorni successivi stesi delle note per avere una registrazione delle ragioni che avevano spinto le varie persone a operare una determinata scelta. Il risultato non era quello che mi ero aspettato; in effetti, i miei amici mi «spiazzarono» notevolmente. Ebbi con loro accanite discussioni su ciò che era «razionale» fare, e tutti loro si mostrarono molto interessati alla faccenda.

Vorrei riportare alcune delle impressioni espresse dai miei amici coinvolti in questa situazione deliziosamente ingannevole. David Policansky iniziò la sua telefonata in modo molto stringato: «O.K., Hofstadter, dammi i 19 dollari!» Poi diede questa spiegazione alla sua scelta della defezione: «Quel che ci chiedi, in effetti, è schiacciare uno fra due bottoni, sapendo soltanto che se schiacciamo il bottone *D* otteniamo più che se schiacciamo il bottone *C*. Quindi *D* è meglio. L'essenza del mio ragionamento è questa. Io defeziono.»

Martin Gardner (sì, ho chiesto anche a lui di partecipare) ha espresso molto vivamente il tormento mentale che lui e molti altri avevano provato. «Orribile dilemma. Davvero non so che cosa fare. Se volessi massimizzare la mia vincita, sceglierei *D* aspettandomi che gli altri facciano altrettanto; per massimizzare la mia soddisfazione sceglierei *C* sperando che anche altri facciano lo stesso (in virtù di un kantiano imperativo categorico). Non so, però, come ci si dovrebbe comportare *razionalmente*. Si finisce in regressi all'infinito: "Se tutti fanno *X*, allora dovrei fare *Y*, ma gli altri lo prevederanno e faranno *Z*, e quindi...". Ci si trova intrappolati in un vortice senza fine. È come il paradosso di Newcomb.» Detto questo, Martin defezionò, sia pure con un sospiro di rimpianto.

Chris Morgan, facendo eco al senso di confusione espresso da Martin, disse: «Più che altro per intuizione, arrivo alla conclusione che non c'è modo di risolvere i paradossi legati a questa situazione. Ho quindi deciso di lanciare una moneta perché non posso prevedere quello che faranno gli altri. Penso - ma non posso esserne certo - che si negheranno tutti l'un l'altro.» Perciò Chris, mentre era al telefono, lanciò una moneta e «scelse» di cooperare.

Sidney Nagel era molto insoddisfatto della sua decisione. «La notte scorsa non sono proprio riuscito a dormire per pensare al problema. *Volevo* cooperare, ma non riuscivo a trovare un modo per giustificarlo. A mio modo di vedere, quello che faccio io non può modificare quello che fanno gli altri. Potrei anche considerare che gli altri abbiamo già deciso, nel qual caso il meglio che possa fare per me è giocare *D*.»

Robert Axelrod, il cui lavoro dimostra la superiorità delle strategie cooperative nella versione *iterata* del dilemma del prigioniero, non vedeva alcuna ragione per cooperare in una partita a colpo secco e defezionò senza rimorsi.

Dorothy Denning fu molto breve: «Mi sembra che, defezionando, ottenga sempre almeno lo stesso risultato che se avessi cooperato. Quindi defeziono.» Fu una delle persone che mi sorpresero. Suo marito, Peter, decise di cooperare. Io avevo previsto il contrario.

Probabilmente avrete tenuto il conto. Finora abbiamo cinque che defezionano e due che cooperano. Supponete di essere al mio posto e che abbiate ricevuto circa un terzo delle telefonate, e che la situazione sia di cinque a due in favore della defezione. Vi azzardereste a estrapolare questa statistica, pronosticando un 14 a sei finale? Quando mai le scelte di sette individui hanno a che fare con le scelte di *altri* 13 individui? Come diceva Nagel, certamente una scelta non può influenzarne un'altra (a meno che crediate in qualche modo alla trasmissione telepatica, un'ipotesi che qui tuttavia non prenderemo in considerazione). Come si potrebbe giustificare un'estrapolazione di questi risultati?

		Giocatore *B*	
		Coopera	Defeziona
Giocatore *A*	Coopera	(3,3)	(0,50)
	Defeziona	(50,0)	(0,01, 0,01)

Una seconda matrice di pagamento per il Dilemma

È chiaro che una simile giustificazione si fonderebbe sull'idea che le persone siano in qualche modo «simili» l'una all'altra. Si fonderebbe sull'idea che, in decisioni complesse e delicate come questa, la gente si affiderà a una serie di ragioni, idee, pregiudizi e vaghe nozioni, alcune delle quali tenderanno a spingere in una direzione, altre in direzione opposta, ma il cui impatto globale porterà una certa percentuale di persone verso un'alternativa e un'altra percentuale verso l'altra. Non si può prevedere in anticipo quali saranno quelle percentuali, ma, dato un campione di persone nella situazione, si può sperare che le loro decisioni siano «tipiche». L'idea, allora, che le prime risposte che danno cinque a due in favore della defezione possano essere estrapolate a un risultato finale di 14 a 6 (circa), si baserebbe sul presupposto che le sette persone agiscano in «modo tipico» per persone sottoposte a queste pressioni mentali in conflitto.

L'intoppo è che le pressioni mentali non sono del tutto esplicite; sono richiamate, ma non completamente messe in chiaro, dalla formulazione della lettera. Ogni persona porta un unico insieme di immagini e associazioni a ogni parola e concetto, ed è l'insieme di quelle immagini e associazioni a creare collettivamente nella mente di quella persona un insieme di spinte mentali simile all'insieme delle sollecitazioni che agiscono sulla crosta terrestre in una zona sismica. Quando una persona prende le sue decisioni, si scopre che incidenza hanno tutte quelle sollecitazioni che agiscono in direzioni differenti, come un insieme di vettori forza variamente orientati, la cui lunghezza è influenzata da fattori segreti o non misurabili. Il presupposto che sia valido estrapolare si basa sull'idea che tutte le persone siano internamente simili e differiscano solo un po' per il peso che attribuiscono a certe nozioni.

In questo modo, la decisione di ciascuna persona può essere paragonata a un esperimento geofisico con l'obiettivo di

prevedere quando e dove si verificherà un terremoto. Si predispone un modello della crosta terrestre e si introducono i dati che rappresentano le migliori conoscenze sulle tensioni interne. Sapete che vi sono ampi margini di incertezza nelle vostre conoscenze e quindi dovete limitarvi a scegliere, per le variabili, valori che vi sembrano «ragionevoli». Nessuna singola esecuzione della vostra simulazione avrà pertanto un forte potere predittivo, ma questo non ha importanza: eseguite la simulazione e ottenete una linea di faglia che vi dice dove si sposta la crosta terrestre simulata. Poi tornate indietro, scegliete altri valori nei domini di quelle variabili e procedete daccapo a un'altra simulazione. Se eseguite la simulazione più volte, alla fine emergerà uno schema che vi rivela dove e come è probabile che la crosta si muova e dove invece si trova roccia solida.

Questo tipo di simulazione dipende da un principio essenziale della statistica: se si fanno assumere alle variabili pochi valori campione scelti a caso dai rispettivi domini, l'esito globale determinato da un grappolo di tali variabili inizierà a emergere dopo pochi tentativi e fornirà presto un modello accurato. Non c'è bisogno di eseguire la simulazione milioni di volte per veder emergere tendenze valide.

Chiaramente, è questo il presupposto su cui si basano le reti televisive americane per predire i risultati delle elezioni nazionali partendo dai primi risultati di poche città scelte negli stati dell'Est. Certo non pensano che la libera scelta sia «più libera» nell'Est che nell'Ovest, e che l'Ovest seguirà le scelte fatte dall'Est, quali che siano. Il fatto è che l'insieme delle pressioni emotive e intellettuali sui votanti è pressappoco identico in tutta la nazione. Ovviamente, nessun individuo può essere preso come rappresentante dell'intera nazione, ma si può presupporre che un gruppo ben scelto di abitanti degli stati costieri dell'Est sia rappresentativo dell'intera nazione per quel che riguarda l'influenza delle varie pressioni elettorali, così che probabilmente le loro scelte indicano le tendenze generali dell'elettorato.

Supponiamo che risulti che la Belknap County nel New Hampshire e la Modoc County in California hanno prodotto, nel corso di molte elezioni nazionali, risultati molto simili. Ne seguirebbe che una delle due contee ha esercitato qualche tipo di influenza causale sull'altra? Doveva esserci qualche strana risonanza cosmica mediata da «magia simpatetica»? Certo no. Basta che gli elettorati delle due contee siano simili; allora le pressioni che determinano il modo in cui la gente vota diventano l'elemento predominante e renderanno automaticamente analoghi i risultati. Non è più misterioso di quanto lo sia l'osservazione che una studentessa della Belknap County e uno studente della Modoc County daranno la stessa risposta se viene loro chiesto di dividere 507 per 13: le leggi dell'aritmetica sono le medesime in tutto il mondo e operano allo stesso modo in menti lontane senza alcun bisogno di una «magia simpatetica».

Tutto questo è elementare buon senso; dovrebbe essere il genere di cose che una persona con una certa istruzione capisce automaticamente. Eppure, da un punto di vista emotivo non si può fare a meno di considerarlo un po' particolare, perché sfida l'idea di libero arbitrio e considera le decisioni delle persone come l'esito di combinazioni di pressioni dai valori sconosciuti. D'altra parte, è forse meglio considerare le decisioni in quel modo piuttosto che attribuirle al libero arbitrio, una nozione filosofica quanto meno oscura.

Quanto sopra può essere sembrata un'inutile digressione sulla statistica e il problema del contrasto fra azioni individuali e prevedibilità del comportamento di gruppo, ma in realtà ha pienamente a che fare con l'«azione corretta» da compiere nel dilemma della mia lettera. C'eravamo posti la seguente domanda: fino a che punto si può prendere il comportamento di *poche* persone come indicazione di ciò che *tutti* faranno? Si può anche andare più in là: fino a che punto il comportamento di *una* persona può essere preso come indicazione di ciò che *tutti* faranno? La versione finale della domanda, formulata in prima persona, ha anche un divertente risvolto: fino a che punto la *mia* scelta mi dice qualcosa sulle scelte degli altri partecipanti?

Si potrebbe pensare che ogni persona è del tutto unica e che quindi non ci si può basare su nessuno per prevedere il comportamento degli altri, soprattutto in una situazione altamente problematica. Ma non è così semplice. Io avevo cercato di progettare la situazione in modo che tutti ne avessero la medesima immagine. Nel centro dell'immagine si supponeva vi fosse la nozione che tutti, in quella situazione, usassero *solo il ragionamento* - compreso il ragionamento sul ragionamento - per arrivare a una risposta.

Se il ragionamento provoca una risposta, tutti avrebbero dovuto arrivare indipendentemente a quella risposta (proprio come la studentessa della Belknap County e lo studente della Modoc County darebbero 39 come risposta alla divisione). Rendersi conto di questo fatto è proprio il passo critico nel ragionamento verso la risposta corretta, ma sfortunatamente è sfuggito a quasi tutti quelli a cui avevo mandato la lettera. (È per questo che vorrei aver incluso nella lettera un paragrafo in cui si sottolineasse la razionalità dei partecipanti.) Una volta che ci si rende conto di questo fatto, comincia ad apparir chiaro che *o* tutti i giocatori razionali scelgono *D o* tutti i giocatori razionali scelgono *C*. Questo in effetti è il punto.

Qualsiasi numero di pensatori razionali ideali, posti di fronte alla stessa situazione e tutti alle prese con gli stessi tormenti dell'agonia del ragionamento, arriveranno alla fine necessariamente alla stessa risposta se il solo ragionamento è la giustificazione definitiva della loro conclusione. In caso contrario, il ragionamento sarebbe soggettivo, non oggettivo come l'aritmetica. Una conclusione raggiunta per ragionamento sarebbe una questione di preferenza, non di necessità. Ora, *qualcuno* può anche pensarla così, ma i pensatori razionali capiscono che un'argomentazione valida deve avere una forza *universale,* altrimenti è semplicemente un'argomentazione non valida.

Se ammettete questo, siete al 90 per cento del cammino. Ora basta chiedersi: «Dato che abbiamo ricevuto tutti la stessa lettera, qual è la scelta più logica? Cioè, qual è il mondo migliore per *il singolo* pensatore razionale: uno in cui tutti cooperano o uno in cui tutti defezionano?» La risposta è immediata: «Ottengo 57 dollari se tutti cooperiamo, 19 dollari se tutti defezioniamo. Chiaramente io preferisco 57 dollari, quindi cooperare è preferibile per questo pensatore razionale. Dato che io sono tipico, cooperare deve essere preferito da *tutti* i pensatori razionali. Quindi coopererò.» Si potrebbe formulare la cosa in modo più «magico»: «Se scegliessi *C*, tutti sceglierebbero *C*, quindi otterrei 57 dollari. Se scegliessi *D*, tutti sceglierebbero *D*, quindi otterrei 19 dollari. Io preferisco 57 dollari a 19 dollari, quindi sceglierò *C*. Allora tutti sceglieranno *C* e io avrò 57 dollari.»

A molti tutto questo può sembrare come credere nel voodoo o nella magia simpatetica, la visione di un universo permeato da tenui fili di sincronicità, che portano pensieri da mente a mente come la posta pneumatica porta messaggi da una parte all'altra di Parigi, e che portano le persone a risuonare secondo una segreta armonia. Niente è più lontano dal vero. Questa soluzione non dipende affatto dalla telepatia o da bizzarre forme di causalità. È solo che l'affermazione «Sceglierò *C* e quindi tutti sceglieranno *C*», pur essendo del tutto corretta, è formulata in modo un po' fuorviante. La parola «scelta» è incompatibile con il carattere coercitivo della logica. Gli studenti non *scelgono* quanto fa 507 diviso 13: lo calcolano. Analogamente, la mia lettera non consentiva scelte: chiedeva un ragionamento. Allora un modo migliore per formulare l'affermazione «voodoo» sarebbe: «Se il ragionamento *mi* guida a dire *C*, allora, visto che non sono differente da nessun altro per quanto riguarda il pensiero razionale, il ragionamento guiderà tutti a dire *C*.»

La corrispondente incursione nel mondo opposto («Se scegliessi *D*, tutti sceglierebbero *D*») può essere capita meglio se la paragoniamo alla riflessione fatta dalla studentessa della Belknap County prima di dividere: Dunque, credo che 13 stia in 507 circa 49 volte - o forse 39. Devo calcolarlo. Ma quel che so per certo fin da adesso è che se scoprirò che il risultato è 49, allora anche il ragazzo della Modoc County scriverà 49, e se troverò che la risposta è 39, la troverà anche lui. Non c'è nessuna trasmissione segreta; basta l'universalità e uniformità dell'aritmetica. Analogamente, l'argomentazione «Qualsiasi cosa faccia, verrà fatta da tutti» è solo l'espressione della fiducia che il ragionamento sia universale, almeno tra i pensatori razionali, non significa abbracciare qualche mistica forma di causalità.

Quell'analisi mostra perché dovreste cooperare anche quando lì sul tavolo davanti a voi stanno le buste opache con le risposte degli altri giocatori. Posti così concretamente di fronte all'inalterabile insieme di *C* e di *D*, potreste pensare: qualsiasi cosa abbiano fatto, è meglio che giochi *D* piuttosto che *C*, perché certo quello che *ora* scelgo non può avere effetto retroattivo su ciò che essi hanno scelto. Quindi defeziono. Questo ragionamento, però, presuppone che la logica che ora vi guida a giocare *D* non abbia connessione o relazione con la logica che ha in precedenza guidato gli altri giocatori alle loro decisioni. Ma se accettate ciò che era affermato nella lettera, allora dovete concludere che la decisione che ora prendete sarà rispecchiata dalle giocate nelle buste che avete di fronte. Se la logica ora *vi* obbliga a giocare *D*, essa ha *già* obbligato gli altri a fare lo stesso, e per le stesse ragioni; per converso, se la logica vi obbliga a giocare *C*, ha già obbligato gli altri a fare allo stesso modo.

Immaginate di avere sul tavolo una pila di buste con le risposte di altre persone al problema aritmetico «Quanto fa 507 diviso 13?». Avendo calcolato un po' frettolosamente, state per mettere nella busta un «49», quando all'ultimo minuto decidete di controllare, scoprite il vostro errore e cambiate il 4 in un 3. Vi immaginate che in quel momento tutte le risposte nelle buste passino improvvisamente da 49 a 39? Naturalmente no. Semplicemente riconoscete che è cambiata la vostra *immagine* del contenuto di quelle buste, non il contenuto stesso. Prima pensavate che ci fossero molti 49, mentre ora pensate che ci siano molti 39. Non ne segue, però, che ci sia stato un istante in cui abbiate pensato: stanno tutti passando da 49 a 39. In effetti, sareste pazzi a crederlo.

Lo stesso vale per i *C* e i *D*. Se dapprima siete inclini a giocare in un certo modo, ma poi una più attenta considerazione vi porta a giocare nell'altro modo, gli altri giocatori non vi seguiranno retroattivamente o sincronicamente. Se, d'altra parte, li accreditate della capacità di vedere la logica che voi avete visto, dovete presupporre che le loro risposte siano uguali alla vostra. In breve, non sarete in grado di «far loro le scarpe»; siete semplicemente in lega con loro, vi piaccia o meno la cosa! O tutti *D* o tutti *C*. Signori, fate la vostra scelta.

In realtà, dire «Fate la vostra scelta» è quanto mai fuorviante. *Non* è che voi possiate semplicemente «scegliere» e poi gli altri - anche nel passato - vi seguano magicamente! Il punto è che se state scegliendo secondo quella che ritenete una *logica* stringente, se veramente rispettate il carattere coercitivo della vostra logica, allora dovete credere che anche gli altri la accetteranno, il che significa che certamente *non* state «scegliendo». In effetti, più siete convinti di ciò che giocate, più dovreste essere certi che anche gli altri giocheranno (o hanno giocato) allo stesso modo e per le stesse ragioni. Questo vale sia che giochiate *C* sia che giochiate *D*, e qui sta il centro della soluzione. Invece di essere un paradosso, è una soluzione che si autorafforza, un circolo benigno della logica.

Se ancora vi sembra di vedere della causalità retroattiva, prendete in considerazione questo raccontino che può servire a illuminare quanto detto. Supponiamo che voi e Giulia amiate la musica classica. Nel corso degli anni avete scoperto d'avere dei gusti incredibilmente simili in musica: una notevole coincidenza. Un giorno venite a sapere che si terranno due concerti contemporaneamente nella città in cui vivete. Entrambi vi sembrano ottimi, ma al Concerto *A* non potete proprio mancare, mentre il Concerto *B* è una forte tentazione a cui potete però resistere. Il Concerto *B*, tuttavia, vi incuriosisce enormemente, perché l'esecutore è Zilenko Buznani, un violinista di cui avete sentito dire meraviglie.

Dapprima siete indispettiti, ma poi un pensiero vi attraversa la mente: forse posso almeno avere da Giulia un resoconto di prima mano sul modo di suonare di Buznani. Dato che io e lei ascoltiamo la musica quasi con le stesse orecchie, il fatto che vada *lei* è quasi come se andassi io. Questo vi conforta un po', ma poi vi accorgete che qualcosa non va. Per le vostre stesse ragioni, Giulia insisterà per ascoltare il Concerto *A*. Dopo tutto, lei ama la musica proprio come voi - ed è per questo che volete che lei vi racconti del Concerto *B*. Più pensate che Giulia abbia i vostri stessi gusti, più volete che vada all'altro concerto. Ma più i suoi gusti sono uguali ai vostri, meno vorrà andarci!

Siete legati dal vincolo dei gusti comuni. E se i vostri gusti sono abbastanza diversi perché non siate d'accordo su quale dei due concerti sia meglio, allora comincereste a perdere interesse in quello che Giulia potrebbe raccontarvi, perché nella sua opinione non avreste più quella fiducia che si ha per qualcuno che ascolta la musica «con le vostre stesse orecchie». In altre parole, sperare che scelga il Concerto *B* è inutile, in quanto mina alla base le ragioni per cui vi *preoccupate* di quale concerto sceglie!

Mi sembra che l'analogia sia chiara. Scegliere *D* mina le ragioni per sceglierlo. Nella misura in cui tutti voi che partecipate al gioco *siete* realmente pensatori razionali, penserete realmente allo stesso modo. Nei miei intenti, la mia lettera doveva stabilire al di là di ogni dubbio che siete tutti «in sincronia», cioè assicurarvi di poter contare sulla razionalità degli altri, che è tutto ciò di cui avete bisogno.

Beh, non proprio tutto. Avete bisogno di contare non solo sulla loro razionalità, ma anche sul fatto che essi possano contare sulla razionalità di tutti gli altri, *e* sul fatto che essi possano contare sul fatto che tutti possono contare sulla razionalità di tutti, e così via. Io chiamo super-razionale un gruppo di persone che ragionano legate l'una all'altra da questa relazione. Dei pensatori razionali, per definizione ricorsiva, inseriscono nei loro calcoli il fatto di trovarsi in un gruppo di pensatori super-razionali. Da questo punto di vista, assomigliano a particelle elementari «rinormalizzate».

Il modo in cui un elettrone rinormalizzato interagisce con, diciamo, un fotone rinormalizzato, tiene conto del fatto che la struttura quantistica del fotone include degli «elettroni virtuali» e che la struttura quantistica dell'elettrone include dei «fotoni virtuali»; inoltre, tiene conto del fatto che tutte queste particelle virtuali (esse stesse rinormalizzate) interagiscono anche l'una con l'altra. Ne segue una cascata infinita di possibilità, di cui, però, la natura tiene conto in un sol colpo. Analogamente, la super-razionalità, cioè il ragionamento rinormalizzato, comporta l'osservazione di tutte le conseguenze del fatto che altri ragionatori rinormalizzati sono coinvolti nella stessa situazione - e questo in un tempo finito, senza soccombere a un infinito regresso di ragionamento sul ragionamento sul ragionamento...

La risposta che speravo di ricevere da tutti era *C*. Non ero così ottimista da credere davvero che tutti arrivassero a questa conclusione, ma mi aspettavo che la maggioranza lo facesse; di qui il mio sgomento quando invece le prime risposte si rivelarono fortemente in favore della defezione.

Altre telefonate portarono qualche *C*, ma per ragioni errate. Dan Dennett cooperò dicendo: «Preferirei essere la persona che compra il ponte di Brooklyn piuttosto che quello che lo vende. Analogamente, mi sentirei meglio a spendere 3 dollari guadagnati cooperando che 10 dollari guadagnati con la defezione.»

Charles Brenner, che mi ero immaginato un sicuro *D*, mi sorprese rispondendo *C*. Quando gli chiesi il perché, mi rispose candidamente: «Perché non voglio finire su una rivista di circolazione mondiale come uno che defeziona.» Molto bene. Sappi, o mondo, che Charles Brenner è uno che coopera.

Molti accarezzarono l'idea che tutti pensassero «circa lo stesso», ma non la presero abbastanza sul serio. Scott Buresh mi confidò: «Non è stata una scelta facile. Mi son trovato a oscillare avanti e indietro. Sono partito da un presupposto, che tutti attraversassero i miei stessi processi mentali. Ho scoperto di voler cooperare circa un terzo del tempo. Sulla base di quel calcolo e del presupposto che fossi tipico, ho calcolato che circa un terzo delle persone avrebbero cooperato. Calcolando quanto avrei realizzato con sei o sette persone che cooperavano, ho scoperto che dicendo *D* avrei preso tre volte di più che dicendo *C*. Quindi dovevo defezionare. L'acqua cerca il suo livello, e io sono sceso nell'angolo basso a destra della matrice.» A quel punto dissi a Scott che fino a quel momento l'ampia maggioranza aveva defezionato ed egli reagì prontamente: «Quei traditori - come possono defezionare tutti? Mi fa una rabbia! Sono veramente deluso dei tuoi amici, Doug.»

Lo ero anch'io, quando ebbi i risultati finali: 14 persone avevano defezionato e

sei avevano cooperato - esattamente quanto avrebbero previsto le reti televisive. Quelli che avevano defezionato ricevettero così 43 dollari e quelli che avevano cooperato ricevettero 15 dollari. Mi chiedo cosa sta dicendo ora Dorothy a Peter. Scommetto che sogghigna osservando, «Te l'avevo detto che era meglio in questo modo». Ahimé, che cosa si può fare con della gente così?

Un aspetto sorprendente della risposta di Scott Buresh è che egli in effetti aveva trattato il proprio cervello come una simulazione del cervello degli altri, portando la simulazione abbastanza avanti per avere un'idea di ciò che una «persona tipica» avrebbe fatto. Questo è nel vero spirito della mia lettera. Stabilite quali potevano essere le percentuali, Scott fece un calcolo a mente fredda per massimizzare il suo profitto, basandosi sul presupposto che i cooperanti fossero sei o sette. Naturalmente, scelse la defezione. In realtà, non aveva molta importanza il numero di coloro che avrebbero deciso di cooperare: qualsiasi calcolo del genere sarebbe sempre risultato favorevole alla defezione. Finché ritenete che la vostra decisione sia *indipendente* dalla decisione degli altri, sceglierete la defezione. Quello di cui Scott non teneva conto era che le persone che calcolano a mente fredda dovrebbero tener conto che le persone che calcolano a mente fredda dovrebbero tener conto che le persone che calcolano a mente fredda dovrebbero tener conto che...

Sembrerebbe terribilmente difficile tenerne conto in modo finito, ma in realtà è la cosa più facile del mondo. Significa solo che quei tenaci pensatori razionali si accorgeranno di essere in una situazione simmetrica, cosicché quello che la ragione detta a uno lo detterà a tutti. Da questo punto in avanti, il processo è semplice. Che cosa è meglio per il singolo se si tratta di una scelta universale: la cooperazione o la defezione? Questo è tutto.

In realtà non è proprio tutto, perché ho spazzato sotto il tappeto una possibilità: che forse gettare un dado potrebbe essere meglio che fare una scelta deterministica. Come Chris Morgan, si potrebbe pensare che la cosa migliore da fare sia scegliere C con probabilità p e D con probabilità $1-p$. Chris poneva arbitrariamente p uguale a 1/2, ma potrebbe essere un qualsiasi numero compreso tra 0 e 1, dove i due estremi rappresentano rispettivamente la defezione e la cooperazione. Quale valore di p sarebbe scelto da giocatori super-razionali? È facile calcolarlo in un Dilemma del prigioniero a due, dove si presuppone che entrambi i giocatori usino lo stesso valore di p. Il guadagno atteso per ciascuno dei due, in funzione di p, risulta essere $1+3p-p^2$, che cresce monotonicamente col crescere di p da 0 a 1. Il valore ottimale di p è quindi 1, il che significa certezza della cooperazione. Quando si hanno più giocatori il calcolo diventa più complesso, ma la risposta non cambia: l'attesa è sempre massima quando p è uguale a 1. Quindi questa impostazione conferma la precedente, che non comportava strategie probabilistiche.

Gettare un dado per stabilire il da farsi non aggiunge nulla di nuovo al Dilemma del prigioniero standard, ma per la versione con matrice modificata che ho dato nel poscritto della mia lettera? Vi lascio studiare la cosa da soli. E a proposito del dilemma di Platonia? Qui due cose sono chiare: (1) se decidete di non mandare un telegramma le vostre probabilità di vincere sono zero; (2) se tutti mandano un telegramma, le vostre probabilità di vincere sono zero. Se ritenete che quello che sceglierete sarà lo stesso di quello che sceglieranno tutti gli altri perché siete tutti super-razionali, nessuna delle due alternative è molto attraente. Con il dado, però, si presenta una nuova possibilità: gettate un dado con probabilità p che venga «bene» e mandate il telegramma se e solo se viene «bene».

Ora immaginate 20 persone che facciano questo e calcolate quale valore di p massimizza la probabilità che esattamente *una* persona abbia successo. Risulta che tale valore è $p = 1/20$ •, più in generale, $p = 1/N$, dove N è il numero dei partecipanti. Quando N si avvicina all'infinito, la probabilità che esattamente una persona abbia successo è $1/e$, meno del 37 per cento. Con 20 giocatori super-razionali che gettino tutti un dado icosaedrico, la probabilità che *voi* risultiate il vincitore è vicina a $1/(20e)$, un po' meno del 2 per cento. Non è poi male! Sempre parecchio meglio dello zero per cento.

L'obiezione sollevata da molte persone è: «E se il mio lancio risulta negativo? Perché non dovrei mandare comunque il mio nome? Dopo tutto, se non lo faccio non avrò alcuna possibilità di vincere. Non starei meglio che se non avessi mai lanciato il dado e avessi rinunciato volontariamente.» Questa obiezione sembra a tutta prima convincente, ma in realtà è erronea, essendo basata su una cattiva interpretazione del significato di «prendere una decisione». Una vera decisione che si conformi al lancio di un dado significa che ci si *deve* conformare al lancio del dado; se in certe circostanze ignorate il dado e fate qualcos'altro, allora non avete mai preso la decisione che avevate detto di aver preso. La vostra decisione è rivelata dalle vostre azioni, non dalle vostre parole prima di agire.

Se vi piace l'idea di lanciare un dado ma temete che la vostra forza di volontà non vi consenta di resistere alla tentazione di defezionare, immaginate un terzo «bottone di Policansky»: su di esso c'è una L (per «lanciare») e, se lo schiacciate, lancia un dado e poi irrevocabilmente manda il vostro nome o non lo manda a seconda del risultato del dado. In questo modo non vi è consentito di ritornare sulla vostra decisione dopo il lancio del dado. Schiacciare quel bottone vuol dire prendere una *vera* decisione che si conformi al lancio di un dado. Per un comune essere umano sarebbe più facile essere protetto dalla tentazione, ma un giocatore super-razionale non avrebbe timore di rimangiarsi la parola dopo un lancio negativo.

[*Spinto dalla riflessione sulle varie versioni del Dilemma del prigioniero, sul problema della cooperazione e su quest'ultima questione del rimangiarsi la parola di fronte a una forte tentazione, Hofstadter ha pensato bene di mettere ulteriormente alla prova le proprie teorie e... i propri lettori, indicendo una «lotteria» aperta a tutti i lettori (e non) di «Scientific American». Il premio in palio è di dollari 1 000 000/N, dove N sarà il numero delle iscrizioni pervenute. In altre parole, se una sola persona avrà inviato la propria iscrizione alla lotteria potrà aggiudicarsi un premio di un milione di dollari; con mille iscritti, il vincitore avrà diritto a un premio di mille dollari, e via di seguito. Da regolamento, una medesima persona può inviare più iscrizioni con una sola cartolina, indicando con un numero intero il numero delle sue iscrizioni. Con 100 iscrizioni, le probabilità di vincere sono maggiori, ma il premio è corrispondentemente minore. Purtroppo le iscrizioni si chiudevano il 30 giugno 1983 e, per il ritardo standard fra la pubblicazione dell'edizione americana di «Scientific American» e la nostra rivista, non abbiamo avuto la possibilità di estendere l'invito alla partecipazione da queste colonne. Ci resta solo da aspettare che Hofstadter vagli i risultati e ci faccia partecipi delle sue idee su questa nuova forma di dilemma. n.d.r.*]